GATE STACK FOR SUB-50nm CMOS DEVICES: MATERIALS, ENGINEERING, AND MODELING

by

Igor Polishchuk

Memorandum No. UCB/ERL M02/19

14 June 2002

GATE STACK FOR SUB-nm CMOS DEVICES: MATERIALS, ENGINEERING, AND MODELING

by

Igor Polishchuk

Memorandum No. UCB/ERL M02/19

14 June 2002

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Gate Stack for Sub-50nm CMOS Devices: Materials, Engineering, and Modeling

by

Igor Polishchuk

B.S. (California Institute of Technology) 1997
M.S. (University of California, Berkeley) 1999

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Chenming Hu, Chair
Professor Tsu-Jae King
Professor Eicke R. Weber

Spring 2002

Gate Stack for Sub-50nm CMOS Devices: Materilas, Engineering, and Modeling

by

Igor Polishchuk

Abstract

Gate stack for sub-50nm CMOS devices: materials, engineering, and modeling

by

Igor Polishchuk

Doctor of Philosophy in Engineering

Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Chenming Hu, Chair

As CMOS technology continues to scale beyond the 100-nm node, many of the materials currently used in CMOS fabrication approach their physical limits. For example, the SiO_2 gate dielectric is now only several molecular layers thick and can no longer serve as a good insulator between the gate and the channel of an MOS transistor. It is expected that new materials such as high-κ dielectrics and metal gate electrodes will have to be used in CMOS fabrication in order to ensure continued scaling of the technology.

High-κ materials have been shown to successfully reduce the tunneling leakage current through the gate dielectrics. However, two important issues related to alternative gate dielectrics still have to be addressed: reliability of these dielectrics and their effect on the mobility of channel carriers. We first examine the reliability of ultra-thin (14 Å equivalent oxide thickness) silicon nitride gate dialectics under both Fowler-Nordheim and hot-carrier stress. The projected lifetime of this dielectric is similar to that of SiO_2 and meets the requirements imposed by device performance. We further attempt to

examine the physical mechanisms responsible for the degradation of ultra-thin silicon nitride through modeling of the random telegraph noise observed in the gate leakage current. We then propose a model that explains the degraded carrier mobility in transistors with alternative gate dielectrics, and suggest some possible ways to preserve the high mobility inherent to the Si-SiO_2 interface. The most straightforward way to preserve high channel-carrier mobility is to include a thin layer of SiO_2 between the channel and the high-κ dielectric layer.

It is therefore likely that future gate dielectrics will be comprised of more than one layer. A simple and precise model for direct tunneling through multi-layer gate dielectrics is indispensable for understanding the scaling of such dielectric stacks. According to the tunneling model proposed here, gate leakage current through various gate dielectrics (both single- and multi-layer) as a function of the equivalent oxide thickness of the dielectric stack is confined to a family of universal lines. Each of the lines is defined by a single number, the tunneling attenuation coefficient, which is a simple function of the dielectric's parameters.

Metal gate electrodes can also help CMOS scaling by eliminating the polysilicon gate depletion. NMOS and PMOS devices for most applications require gate electrodes with two different work functions. While several promising candidates have been proposed for either NMOS or PMOS devices, the integration of the metals with two different work functions on a single CMOS wafer remains a critical challenge. Here we propose an attractive way of making dual-work-function metal gate CMOS transistors based on metal interdiffusion. To demonstrate this proposed CMOS process we fabricated metal -interdiffusion-gate FETs with nickel and titanium gates.

To my wife

Table of Contents

Acknowledgements

I would like to offer thanks to several people for their support during my time in Berkeley.

Professor Chenming Hu was truly inspirational, and provided highly thoughtful counsel and guidance every step of the way. I thank him for believing in me, and more importantly, for teaching me to believe in myself. I am also grateful to have had the chance to learn from Professor Tsu-Jae King. Her vast knowledge and sincere advice, coupled with her dedication to students has been invaluable.

Thanks to Professor Eicke Weber for serving on my qualifying exam and dissertation committees, and for his genuine interest in my work. Thanks also to Professor Jan Rabaey for serving on my qualifying exam committee and for introducing me to the fascinating world of digital design.

I would like to acknowledge Judy Fong, under whose legendary oversight generations of students have flourished. Dr. George Brown was an exceptional mentor during my time at Sematech and beyond.

Special thanks to Dr. Hideki Takeuchi for his generous help and expert advice. I am happily in his debt. I am also grateful to the device group students who came before me, Wen-Chin Lee, Nick Lindert, Dongun Park, Ya-Chin King, Michael Orshansky, Stephen

Tang, Yang-Kyu Choi, Yeh Tung, Dennis Sylvester, Jakub Kedzierski, Xiaodong Jin, Ski Iyer, all of whom have helped me tremendously by sharing their knowledge and experience. Truly, every new student in this group stands on the shoulders of giants.

I would like to warmly acknowledge my peers Pushkar Ranade, Kevin Yang, Yee-Chia Yeo, Qiang Lu, and Leland Chang for their collaborative support and camaraderie. Their presence and input enlivened my entire graduate school experience; I recall our late nights and many discussions with an affection which will only deepen with time. It was also my great pleasure to work with Charles Kuo, Sriram Balasubramanian, Mark Cao, Kevin Cao, Daewon Ha, Xuejue Huang, Qinq Ji, Min She, Peiqi (Patrick) Xuan, Ron Lin, John McHale.

This PhD is as much my family's as it is mine. My family has played a critical role at every turn by continuously instilling in me their infectious enthusiasm and thirst for knowledge.

Finally, there is one person whose contribution was and is truly indispensable - my wife. Eileen's unbounded love, support, and encouragement have helped me face the most difficult of challenges. I am humbly and eternally grateful for the chance to share my life with such a wonderful person.

Chapter 1 Introduction

1.1 Why CMOS device scaling is important

Complimentary metal-oxide-semiconductor (CMOS) technology has been the backbone of the semiconductor industry over the past 4 decades and has, to a large degree, been responsible for the industry's stunning success. The semiconductor industry grew at a remarkable annual rate of 17% between 1956 and 1996. (For comparison, the average rate of growth of the gross world product was only 8% over the same period of time.) Two factors are necessary for a technology to be that successful. First, it has to provide rapid improvements in product performance. Second, the cost of new products has to be kept low to widen the pool of potential customers. Scaling, or the reduction in size, of CMOS transistors has helped to meet both of these goals.

Reducing the transistor gate length increases the amount of current supplied by a transistor. Higher current allows the circuits to switch more quickly, leading to faster computations. This in turn constantly expands the realm of possible applications for semiconductor products.

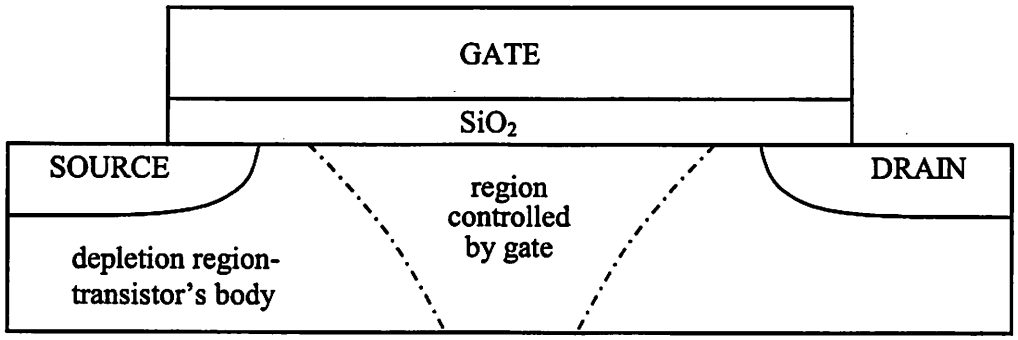
The reduction in transistor size also allows more transistors to be integrated on a single chip. Consequently the complexity and functionality of integrated circuits can be increased while keeping the cost of circuit fabrication low.

1.2 Why CMOS device scaling becomes difficult

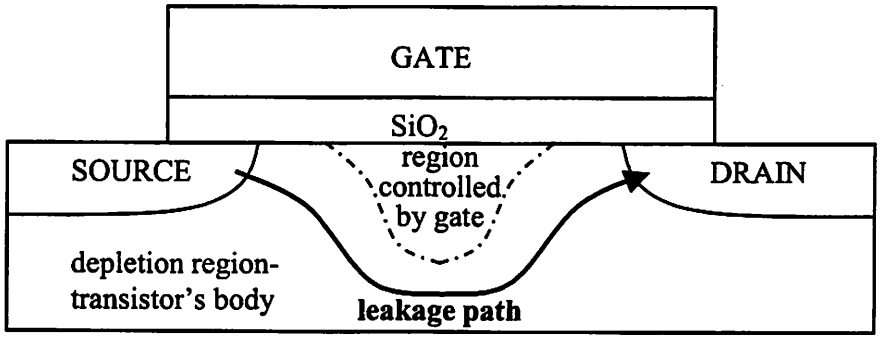
The basic function of a field-effect transistor in digital circuits is to act as a switch (Fig 1.1), i.e. to allow the electric current to flow from the source to the drain when a high voltage is applied to the gate, and to block that current when a low voltage is applied to the gate. The ability to conduct current depends on the electric potential in the body of a transistor. This potential is in turn determined by the amount of capacitive coupling between the transistor's body and various electrodes. It therefore becomes clear that the gate's ability to prevent current from flowing depends on its ability to control the electric potential in the body through capacitive coupling.

In a long channel device (Fig. 1.1a) the gate has exclusive control over the middle region of the body, and can therefore effectively cut off the conduction. However, if the device length is reduced (Fig 1.1b), the regions where the electric potential is controlled by the source and drain electrodes eventually merge and the gate electrode will no longer be able to successfully block the current conduction.

In order to improve gate control the thickness of the SiO_2 dielectric should be reduced (Table 1.1). However if that thickness reduced to less than about 15Å, the tunneling leakage current would increase beyond the allowed limits. This arguably presents the most critical challenge to future CMOS scaling.



(a)



(b)

Fig 1.1. In a long channel transistor (a) the potential induced by the gate electrode can successfully block source-to-drain current flow. When the channel length is reduced, the drain electrode can also influence the electric potential in the middle of the channel, so that a current conduction path can remain open regardless of gate potential.

Table 1.1. International Semiconductor Technology Roadmap [1.1]

predicts rapid reduction of the equivalent gate oxide thickness.

Year of Production	2002	2004	2007	2010	2013	2016
Generation Node (nm)	**130**	**100**	**70**	**50**	**35**	**25**
Physical L_G (nm)	53	37	25	18	13	9
High performance applications						
Equivalent oxide thickness T_{OX} (Å)	12–15	9–14	**6–11**	**5-8**	**4-6**	**4-5**
Gate leakage current (A/cm^2)	60	300	**1500**	**15000**	**50000**	**10^5**
Low power applications						
Equivalent oxide thickness T_{OX} (Å)	22–26	18–22	**12–16**	**9-13**	**8-12**	**7-11**
Gate leakage current (A/cm^2)	2	3	**4**	**15**	**50**	**100**

Solutions exist
Solutions are known
Solutions are NOT known

1.3 How these difficulties can be overcome

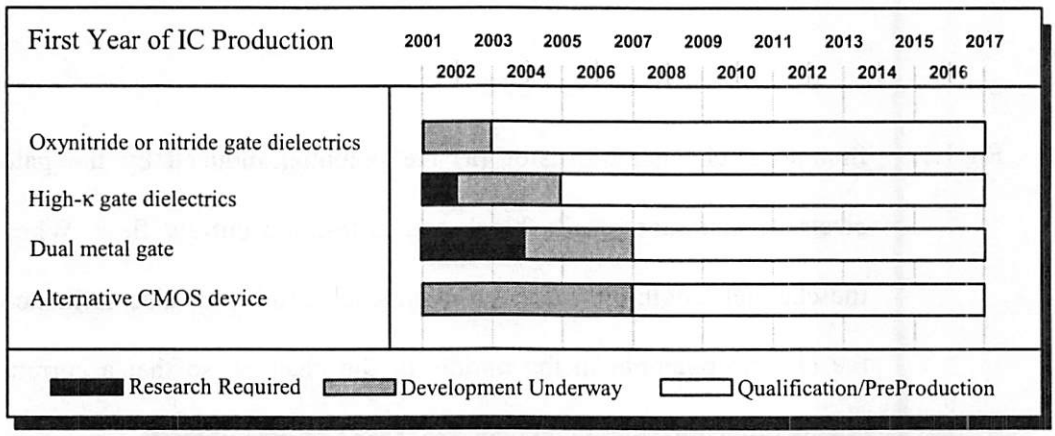


Fig. 1.2. Timeline for research, development, and introduction of new materials and device structures.

1.3.1 New dielectric materials

The reduction of gate oxide thickness is certainly one way to increase the capacitive coupling between the gate electrode and the transistor's body. Unfortunately, it will necessarily entail an unacceptable increase in the gate leakage current. Another way to improve gate-to-body coupling is to increase the permittivity (κ) of the gate dielectric.

A large number of alternate gate dielectrics have been considered with dielectric constants ranging from 7.8 for Si_3N_4 [1.2] to 80 for TiO_2 [1.3]. While the dielectrics with higher κ provide larger reduction in the gate leakage current, they are often very hard to integrate into a CMOS process. Many of these materials are not thermally stable in contact with silicon. In addition, all the dielectrics with $\kappa > 8$ considered so far are metal compounds; introducing a metal directly next to the silicon channel can potentially jeopardize channel carrier mobility and lifetime. Si_3N_4 on the other hand, can be easily integrated into a CMOS process. Despite its relatively low dielectric constant, Si_3N_4 can still meet the gate leakage requirements for high performance CMOS circuits for at least another 15 years. In addition to meeting the performance requirements, Si_3N_4 also has to be electrically reliable as a gate dielectric. Relative to the amount of reliability research done on SiO_2, the reliability of Si_3N_4 is still largely unexplored. In Chapter 2, we shall study the reliability of 14Å equivalent-oxide-thickness Si_3N_4 film under Fowler-Nordheim and hot carrier stress.

While Si_3N_4 will likely suffice as a gate dielectric for high performance applications, a real high-κ dielectric will be needed for low power application in the near future. Besides the fabrication issues already mentioned, it has been observed that high-κ

dielectrics adversely affect channel carrier mobility. We shall address the modeling of channel mobility, especially as it pertains to high-κ dielectrics in Chapter 4.

One of the ways to retain high channel mobility of a Si-SiO_2 system is to introduce a thin interfacial layer of silicon dioxide between the transistor's channel and the high-κ gate dielectric. One should expect that an SiO_2-high-κ dielectric stack is going to have higher leakage current than a pure high-κ gate dielectric of the same equivalent-oxide-thickness. In order to evaluate the amount of leakage current through SiO_2-high-κ dielectric stacks, and predict the scaling limits of such stacks, a simple and precise analytical model is needed. We develop such a model in Chapter 3.

1.3.2 New device structures

Figure 1.1b shows that even in short channel devices, the gate electrode is able to prevent electric current from flowing close to the Si-SiO_2 interface. The only remaining leakage path lies deep in the transistor's body. This suggests an alternative way to improve the off-state source-to-drain leakage current: Reduce the thickness of the transistor's body until this leakage path is cut off. New device structures such as Ultra-Thin-Body (UTB) transistors [1.4] and double-gate transistors [1.5] have been proposed to accomplish this goal.

The most challenging problem for these devices is adjusting their threshold voltages (V_T's). A traditional way of V_T-control for bulk devices is to use substrate implants. In the devices with ultra-thin bodies however, a very high dopant concentration is needed to control the threshold voltage. Unfortunately, high dopant concentration makes V_T very sensitive to variations in body thickness [1.6]. It also may result in

significant mobility degradation. An alternative way to control V_T is to adjust the work function of the gate electrode.

1.3.3 Gate work function engineering

Replacing polysilicon gate with another semiconductor material such as SiGe [1.7], can help control the threshold voltage of MOS devices. By varying the germanium content the work function of a P-type gate can be changed. An N-type SiGe gate however, cannot be used to adjust the threshold voltage of an NMOS device. By using a gate stack comprised of a very thin polysilicon layer and a mid-gap work function metal one can in theory achieve a wide range of V_T-control for both NMOS and PMOS devices [1.8]. However as long as a semiconductor remains at the interface with a gate dielectric, the gate electrode will suffer from the depletion effect at high gate bias. This polysilicon depletion effect (PDE) can substantially reduce the current drive for transistors with thin gate dielectrics [1.9]. The most effective way to eliminate PDE is to switch to a metal gate entirely.

1.3.4 Metal gate work function engineering

Changing work function (i.e. changing the position of the Fermi level) of a semiconductor is relatively easy. Because semiconductor's intrinsic Fermi level lies within the band gap, it can be easily shifted either upwards or downwards by introducing either donor or acceptor atoms. Since a metal does not have a band gap, its work function

can only be changed by either a chemical reaction with another element or by inducing a change in the crystalline phase or crystalline orientation.

One demonstrated way to modify the work function of a metal is through ion implantation. This was first demonstrated by nitrogen implantation into TiN_X [1.10]. By changing the stoichiometry of titanium nitride, its work function was changed by only 0.1 eV. A more significant result of nitrogen implantation has been demonstrated for molybdenum gate electrode [1.11], where the work function was modified from 5.1 to 4.0 eV.

An alternative approach to creating metal gates with varying work functions will be demonstrated in Chapter 5. This approach relies on the interdiffusion of two metals (one with a low and the other with a high work function). The advantage of this approach is that potential damage from ion implantation to the gate dielectric is avoided.

1.4 References

1.1 The International Technology Roadmap for Semiconductors, Semiconductor Industry Association, San Jose, CA, 2001.

1.2 R.P.S. Thakur, A. Ahmad, A. Ditali, A. Martin, A. Kermani, C. Galewski, R. McIntosh, K. Legler, "Ultrathin gate and capacitor dielectric formation using single-wafer processing," *Rapid Thermal and Integrated Processing II*, pp. 401-406, San Francisco, CA, 1993.

1.3 Xin Guo, T. P. Ma, T. Tamagawa, B. L. Halpern, "High quality ultra-thin TiO_2/Si_3N_4 gate dielectric for giga scale MOS technology," *in Tech. Dig. Int. Electron Devices Meeting*, pp. 377-380, San Francisco, CA, Dec. 1998.

1.4 B. Yu, Y.-T. Tung, S. Tang, E. Hui, T.-J. King, C. Hu, "Ultra-thin-body silicon-on insulator MOSFET's for Terabit-scale integration," *Int. Semiconductor Device Research Symp.*, pp. 623-624, 1997.

1.5 D Hisamoto, T. Kaga, Y. Kawamoto, E. Takeda, "A fully depleted lean-channel transistor (DELTA)-a novel vertical ultrathin SOI MOSFET," *IEEE Electron Device Letters*, Vol. 11, No.1, pp. 36-38, Jan. 1990.

1.6 G. G. Shahidi, *et al.*, "A room temperature 0.1 μm CMOS on SOI," *Proc. Dig. Papers, Symp. VLSI Technol.*, pp.27-28, 1993.

1.7 W.-C. Lee, Y.-C. King, T.-J. King, C. Hu, "Investigation of poly-$Si_{1-x}Ge_x$ for dual-gate CMOS Technology," *IEEE Electron Device Letters* Vol. 19, No. 7, pp.247-249, July 1998.

1.8 I. Polishchuk, C. Hu, "Polycrystalline silicon/metal stacked gate for threshold voltage control in metal-oxide-semiconductor field-effect transistors," *Applied Physics Letters*, Vol.76, No.14, pp. 1938-40, 3 April 2000.

1.9 K. F. Schuegraf, C. C. King and C. Hu, "Impact of polysilicon depletion in thin oxide MOS technology," *VLSI TSA*, pp. 86-90, 1993

1.10 H. Wakabayashi, Y. Saito, K. Takeuchi, T. Mogami, T. Kunio, "A novel W/TiNx metal gate CMOS technology using nitrogen-concentration-controlled TiNx film," *in Tech. Dig. Int. Electron Devices Meeting*, pp. 253-256, Washington, DC, Dec. 1999.

1.11 P. Ranade, R. Lin, Q. Lu, Y. -C. Yeo, P. Ranade, H. Takeuchi, T. -J. King, C. Hu, "Molybdenum Gate Electrode Technology for Deep Sub-micron CMOS Generations", *Materials Research Society Symposium Proceedings*, Vol. 670, K5.2.1, Apr. 2001.

Chapter 2 Reliability of Si_3N_4 Gate Dielectric

2.1 Gate dielectric scaling and new reliability concerns

Traditionally the reliability of gate SiO_2 has been represented by the time-to-breakdown, i.e. the time the dielectric can withstand certain electrical stress before breaking down. This dielectric breakdown (now usually referred to as hard dielectric breakdown) is characterized by a sudden change in the conductance of the dielectric by many orders of magnitude. In the example shown in Figure 2.1 the change in conduction of the gate dielectric is at least 8 orders of magnitude. A highly conductive path (a short) is formed through the gate dielectric during a hard breakdown event resulting in a catastrophic failure of an MOS transistor.

As the thickness of the gate dielectric was reduced, it became comparable to the typical size of the stressed-induced defects in the dielectric. This means that a single trap can now influence the conductance of the gate dielectric (even if it does not create a short). The conductance of the dielectric can vary depending on the position of the defect and how the defect aligns and interacts with other defects. Depending on how large the conductance of the dielectric after stress is, this new "breakdown" phenomenon is

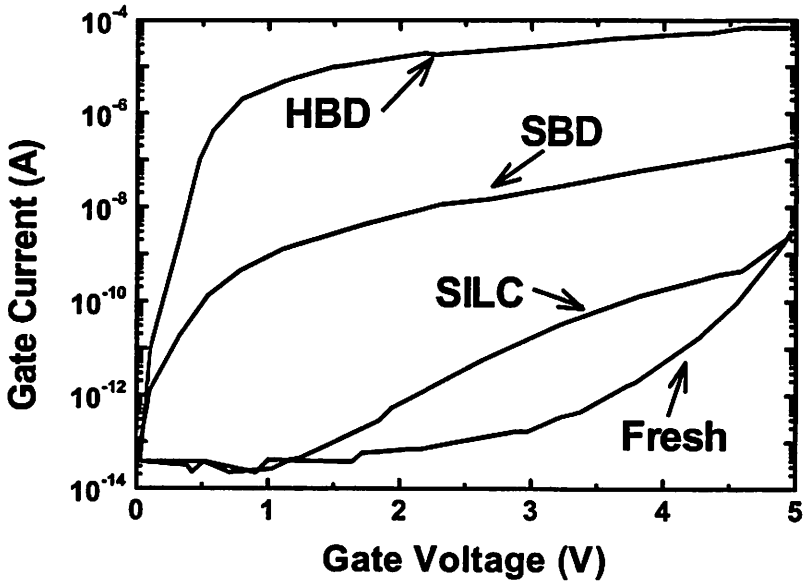


Fig. 2.1. The I_G-V_G characteristics of 4.5 nm oxide capacitors before (fresh) and after stress. Different modes of failure are indicated. (Adapted from [2.1]).

referred to as either soft breakdown (SBD) or stress-induced leakage current (SILC). Soft breakdown is also often associated with the current fluctuations (or noisy signal), as can be seen in Figures 2.2 and 2.3. While many researchers still believe that soft breakdown and SILC are two independent phenomena, we are going to demonstrate (in section 2.5) that SBD and SILC are closely linked.

With continued scaling of CMOS technology not only the thickness of the gate dielectric, but also the supply voltage and the area of the gate are being reduced. All this factors contribute to the fact that soft breakdown becomes increasingly more prevalent compared to hard breakdown [2.3]. It therefore becomes important to pay more attention to these effects and to study the gradual device degradation.

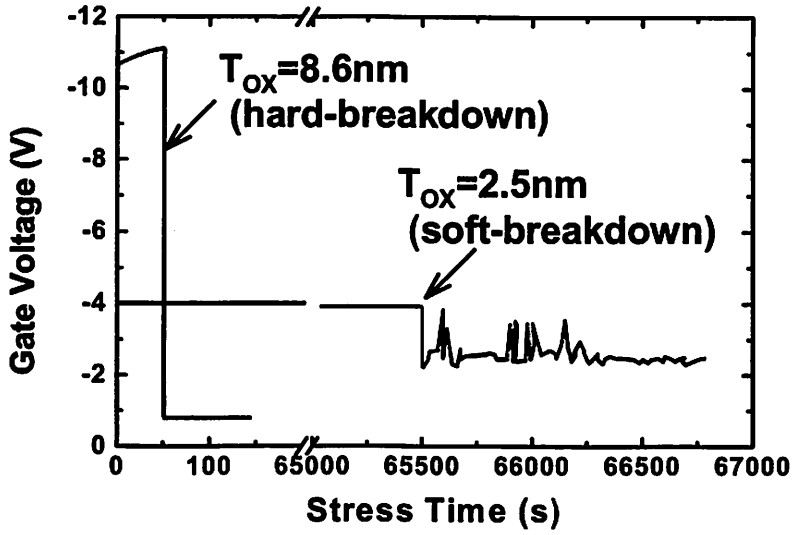


Fig. 2.2. Stress voltage evolution under constant-current stress (I_G = 0.2 A/cm^2 Area = 20×20 μm^2) Adapted from reference [2.2].

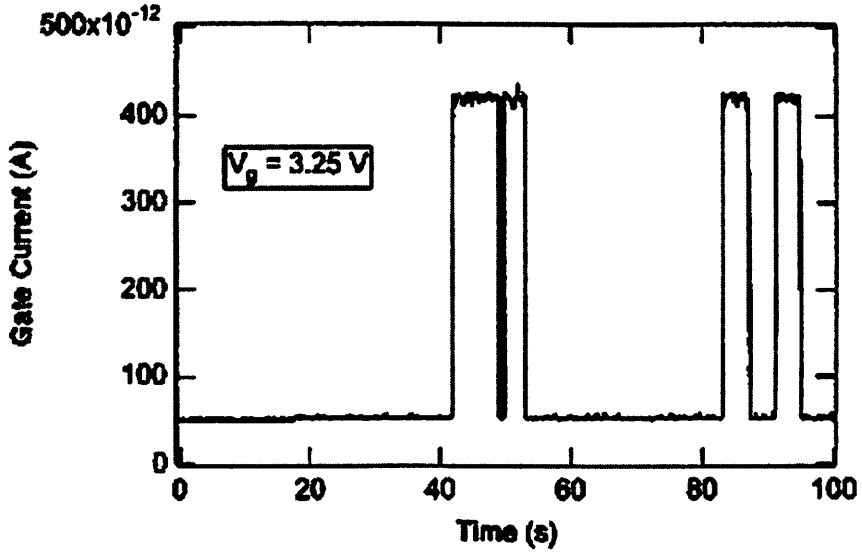


Fig. 2.3. Random telegraph noise observed in the gate current of a capacitor after SBD [2.1].

Even though a device that has experienced a soft breakdown usually remains functional, the soft breakdown effects should not be discounted as a potential reliability problem for the semiconductor product as a whole. For example, a 10-fold increase in the gate leakage current can translate into a 10-fold increase in stand-by power consumption. This can shorten the battery lifetime in a portable semiconductor product from a week to less then a day, which from the customers' point of view can be equivalent to device failure. Thus, it is important to develop the means to analyze and predict the amount of drift in the device parameters under Fowler-Nordheim (FN) stress.

During MOSFET operation, the gate dielectric can also be degraded by hot-carrier stress. A study of hot-carrier reliability for PMOSFETs becomes especially interesting in light of recent publications. First, it has been reported that the degree of hot-carrier degradation in PMOSFETs is approaching that of NMOSFETs for deep-submicron devices [2.4]. Second, the mechanism of the PMOS degradation is changing from hot-electron injection to creation of the dielectric interface states [2.5].

In addition to the new degradation phenomena and mechanisms just described, there is a pressing need to replace SiO_2 with a higher permittivity gate dielectric. The equivalent oxide thickness of the gate dielectric in CMOS integrated circuits is expected to scale below 1.5 nm for most applications by the year 2005 [1.1]. The use of SiO_2 that is thinner than 1.5 nm will not be feasible for some applications due to the high dielectric leakage current. The reliability requirements for the gate dielectric may impose an even stricter limit on the oxide scaling [2.7, 2.8]. Several materials with higher dielectric constants are being investigated as possible replacements for SiO_2. High-κ materials such as Ta, Hf and Zr oxides have recently attracted a lot of attention [2.9-2.11]. However, the

instability of these materials during high temperature processing steps remains a major challenge. Possible metal contamination is another point of concern for metal oxides. Silicon nitride on the other hand, can be easily integrated into CMOS process, and is likely to become the first material to be used as an alternative gate dielectric. The reliability studies presented in this chapter will therefore, focus on silicon nitride gate dielectric.

2.2 Device fabrication and device parameter extraction

2.2.1 Device fabrication

The Si_3N_4 transistors with 100 nm channel length were made using a LOCOS-isolation CMOS process without halo implant. The silicon-nitride gate dielectric was deposited by Jet Vapor Deposition [2.12] at Yale University. Following a 5-minute 800°C anneal in N_2, undoped poly-Si was deposited by LPCVD. I-line lithography and photoresist ashing in O_2 plasma were used to define gate electrodes down to 100 nm. Following the gate patterning, source and drain regions were formed by ion implantation. A "reverse LDD" process [2.13] was used. (In this process, deep source drain regions are formed before LDD extensions, and therefore the LDD extensions undergo less diffusion. Such a process helps to reduce the junction depth and results in improved short channel behavior.) A dose of 1×10^{14} cm^{-2} BF_2^+ was used to form the source and drain extensions in PMOSFETs. Dopants were activated by rapid thermal annealing in N_2.

In the hot-carrier reliability study, we shall also rely on transistors with ultra-thin gate SiO_2 for comparison purposes. These SiO_2 devices were fabricated using a similar process.

The composition of the JVD nitride dielectrics has been previously studied by Auger depth profile analysis [2.12]. The results of this analysis show that the composition of the deposited film is uniform, without a distinct oxygen-reach interfacial layer. The relative atomic concentrations of Si, N, and O in the film are equal to 41%, 46%, and 13% respectively. These numbers indicate that all the Si bonds are satisfied by either N or O, with 84% of those bonds being satisfied by the nitrogen atoms. While JVD material at hand is not technically a stoichiometric Si_3N_4, its low oxygen content clearly sets it aside from a class of gate dielectrics known as oxynitrides. Keeping this caveat in mind, we shall hereafter refer to the dielectric material as nitride or Si_3N_4.

Only P-type devices were successfully fabricated in this CMOS process. Hence the reliability studies in this chapter will exclusively focus on PMOSFETs.

2.2.2 Device parameter extraction

The gate-dielectric equivalent oxide thickness t_{oxeq} was extracted using a quantum mechanical simulator [2.14]. For the nitride transistors, t_{oxeq} is 14 Å and n-well doping concentration is 4×10^{18} cm^{-3} (Fig. 2.4). For the oxide transistors, t_{oxeq} is 16 Å and n-well doping concentration is 7×10^{17} cm^{-3}. Figure 2.5 shows the output current characteristics for both types of short-channel transistors. At low drain bias, the oxide transistor has a higher drain current. This is due to the well-known fact that nitride transistors have lower

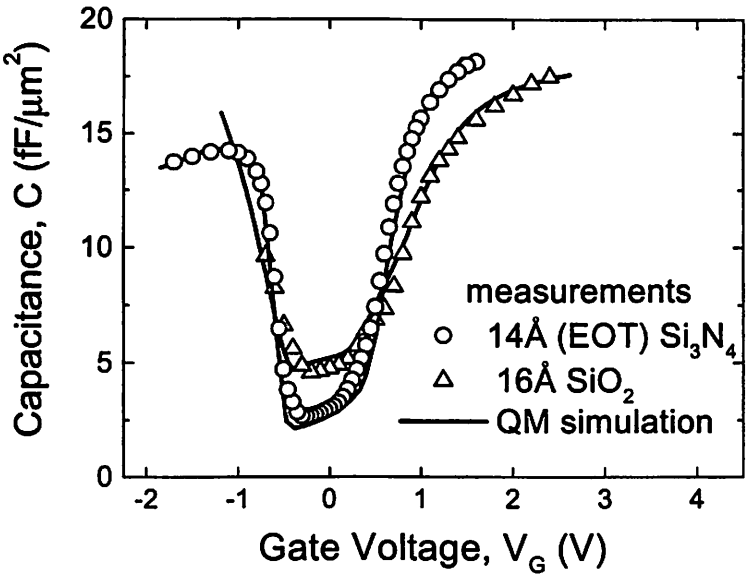


Fig. 2.4. Equivalent oxide thickness is extracted from *C-V* characteristics using a quantum mechanical simulator.

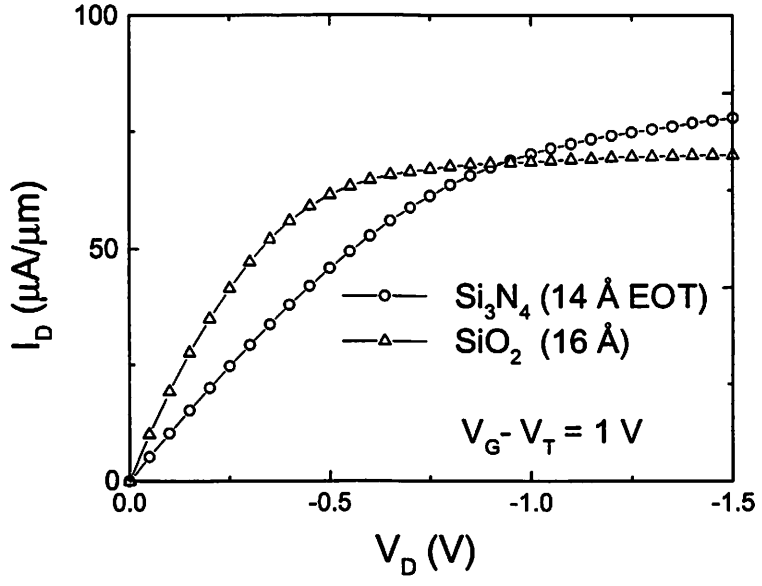


Fig. 2.5. Silicon-nitride PMOSFETs have lower mobility, but higher current drive (at V_{DD}) than silicon-oxide PMOSFETs. $L/W = 0.1 / 1$ μm.

channel mobility. At high drain bias (close to the supply voltage), the situation is reversed, and the nitride transistor has a higher current drive. This higher current drive can be explained by higher inversion charge density in 14 Å Si_3N_4 transistors. The major advantage of the Si_3N_4 gate dielectric is the reduction of the gate leakage current (Fig 2.6). The leakage current of the 14 Å Si_3N_4 PMOSFET is an order of magnitude lower than that of the 16 Å SiO_2 PMOSFET. The reduction in the gate leakage current for NMOSFETs is even larger [2.15].

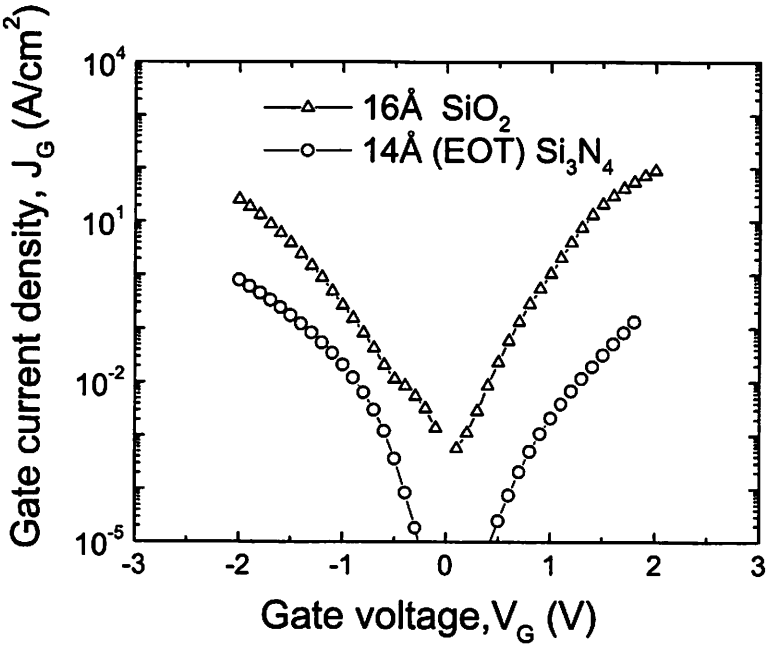


Fig. 2.6. Even though Si_3N_4's equivalent oxide thickness is less than that of SiO_2 by 2 Å, it provides over an order of magnitude reduction in gate leakage current.

2.3 Fowler-Nordheim stress

2.3.1 Background and introduction

Transistors with nitride gate dielectrics as thin as 1.4 nm (equivalent oxide thickness) and channel length of 80 nm have been successfully fabricated [2.16]. Good performance of these devices has been reported. However, it is still necessary to demonstrate good reliability of these devices, before Si_3N_4 can be confidently put forward as a replacement for SiO_2 in the gate stack. Gate nitride reliability research has so far been primarily focused on examining the time-dependent dielectric breakdown (TDDB) [2.17, 2.18]. The consensus has been that TDDB does not present a serious concern for the reliability of the thin nitrides. It has also been noted that the reliability of the thin oxides is limited not by the TDDB but TDDW (the time-dependent dielectric wearout) effects, such as gradual increase in the gate leakage current [2.19, 2.20]. One can expect that TDDW will also play a significant role in thin nitrides as well. This illustrates the need to investigate possible TDDW effects, such as SILC, V_T shift, mobility and subthreshold swing degradation for Si_3N_4.

It has been observed that the degradation of the SiO_2 transistor characteristic due to both hot-electron stress [2.21] and Fowler-Nordheim stress (FN-stress) [2.22] typically follows a power-law dependence on time. In our study, we found that the degradation in the gate leakage current I_G, threshold voltage V_T, subthreshold swing S, and mobility in the transistors with JVD nitride follows the same kind of power-law dependence. Moreover, we propose a set of empirical models that describe the dependence of the device degradation rate on the FN-stress voltage. These models allow us to predict the reliability of the nitride gate dielectric over a 10-year period under low operating

voltages. We shall also examine the possible physical origins of observed power-law dependence of the transistor parameters' drift.

2.3.2 Experiment

Constant voltage FN-stress has been used in this study to predict the reliability of gate nitride under low operating voltages. A stress voltage ranging from –2.7 V to –4 V was applied to the gate of 10×10 μm PMOSFETs, with the source, drain and substrate grounded. This type of stressing should result in the realistic predictions of the gate dielectric reliability. This is because transistors in the integrated circuits are subjected to constant voltage stress rather than constant current stress, and are typically biased into inversion rather than accumulation. The substrate series resistance is approximately 1kΩ, and does not affect the results of the stress measurement as long as the magnitude of the stress voltage remains below 4 V. (The stress current at V_G = -4 V is 50 μA which corresponds to a 50 mV voltage drop across this series resistance.) Constant voltage stress was interrupted from time to time to allow the monitoring of the changes in the gate leakage current I_G, threshold voltage V_T, subthreshold swing S, and transconductance g_m as functions of stress time. Stress-induced leakage current (SILC) is defined here as the difference between the values of I_G for a "stressed" and a "fresh" transistor. The value of SILC at V_G = 1 V was used to quantify the increase in gate leakage current. The values of V_T, S and g_m were obtained from the drain current versus gate voltage (I_D-V_G) characteristics taken at a low drain bias (V_{DS} = - 50 mV). Transconductance values measured at V_G-V_T = - 0.2 V were used to monitor the deterioration of the low field mobility due to stress-generated interface traps.

2.3.3 Estimation of hard-breakdown-limited lifetime

We shall first illustrate that our 14 Å Si_3N_4 gate dielectric is able to withstand sufficiently high fluence of stress charge to ensure that it is not going to experience a hard breakdown event during a 10 year period under normal device operation conditions. Figure 2.7 shows the evolution of the gate leakage current under constant-voltage stress. Stress at low gate biases of –2.7 V and –3 V leads to a very modest increase in the gate current, i.e. SILC. Stress at higher gate biases of –3.3 V and –3.6 V leads to a more dramatic increase in the gate leakage current, i.e. soft breakdown. In practice however, we do not detect any abrupt changes in the dielectric conductance and therefore are not able to distinguish clearly between SILC and soft breakdown. Furthermore, we have not observed hard dielectric breakdown (characterized by an abrupt jump in conduction by a few orders of magnitude) in any of the samples stressed at gate biases as high as –4 V. The bottom curve in Figure 2.7 shows no appreciable increase in the gate leakage current at –2.7 V for a total stress-charge fluence of 2.5×10^6 C/cm^2. The fluence of 2.5×10^6 C/cm^2 corresponds to 10 years of operation at a constant gate bias of 1.2 V and I_G = 7×10^{-3} A/cm^2. It is commonly known that the charge-to-breakdown, Q_{BD}, is much higher at low operating voltages. Thus the intrinsic time to breakdown in excess of 10 years is expected for the 1.4nm gate nitride. This observation is in agreement with the analogous conclusions drawn for 1.9 nm JVD nitride [2.23].

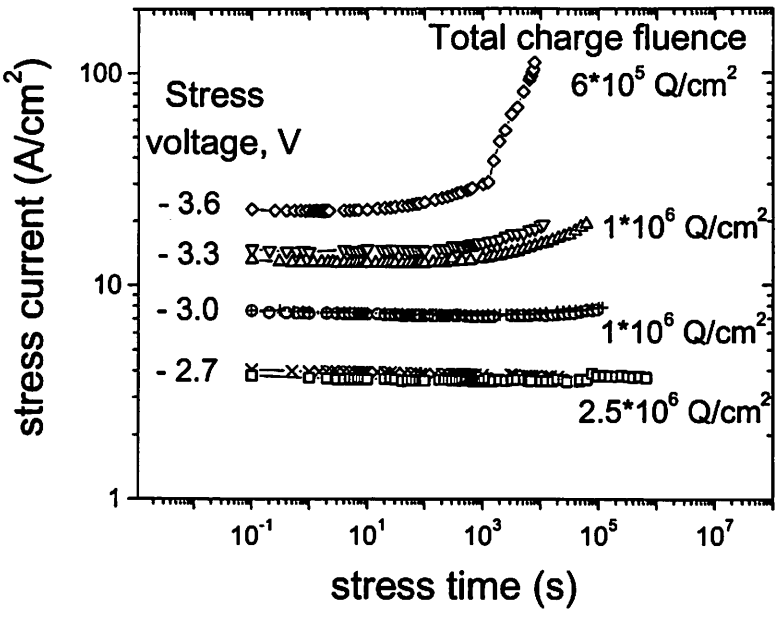


Fig. 2.7. Gate current evolution under constant-voltage stress. Stress voltages vary from –2.7 V to –4 V. The total charge fluence for each of the devices is indicated on the right-hand side. Even though the leakage current increases under higher gate bias stress, no hard dielectric breakdown is detected.

2.3.4 MOSFET parameter drift and lifetime projections

Now we shall turn to the study of degradation in MOSFET parameters under constant-voltage Fowler-Nordhiem stress. Figure 2.8 shows that I_G, $|V_T|$, and S all increase while g_m decreases as the result of FN-stress. While the circuits are not expected to fail due to the gate nitride breakdown, the degradation in the device performance needs to be addressed.

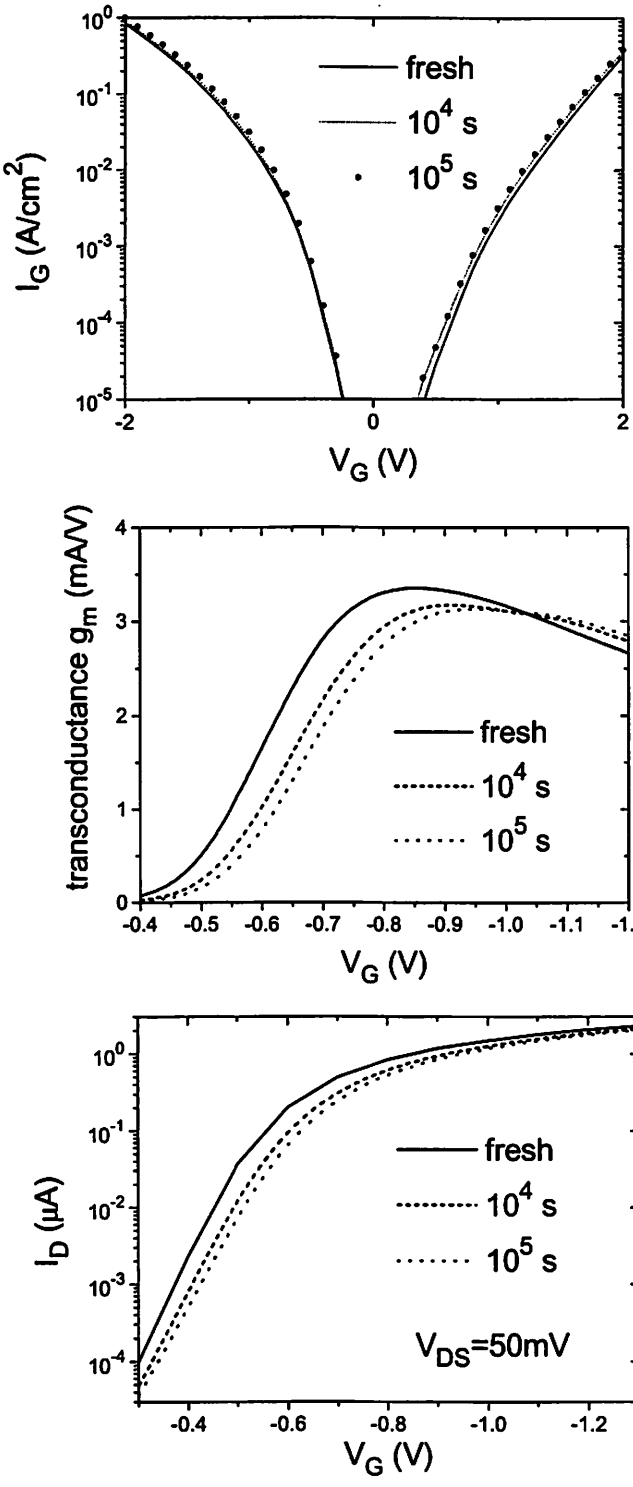


Fig. 2.8. Comparison between **(a)** gate leakage current, **(b)** transconductance, **(c)** I_D-V_G characteristics for fresh and stressed transistors.

The degradation of the aforementioned parameters is commonly attributed to the creation of various types of traps in the gate dielectric. Thus one might expect the changes in one of the parameters would be correlated to the changes in the other parameters. The changes in I_G, V_T, S, and g_m do indeed follow the same power-law dependence as shown in Figure 2.9. Furthermore, the same power-law dependence applies to the results obtained under both high (-3.6 V) and low (-2.7 V) stress voltages. This observation suggests that the following empirical model can be put forward. It would predict the degradation in MOSFET parameters at a low operating voltage as a function of time.

$$\Delta Y = \alpha_Y(V_S)t^{\beta} \tag{2.1}$$

Here Y denotes a MOSFET parameter (such as V_T, for example). β=0.3 is observed in

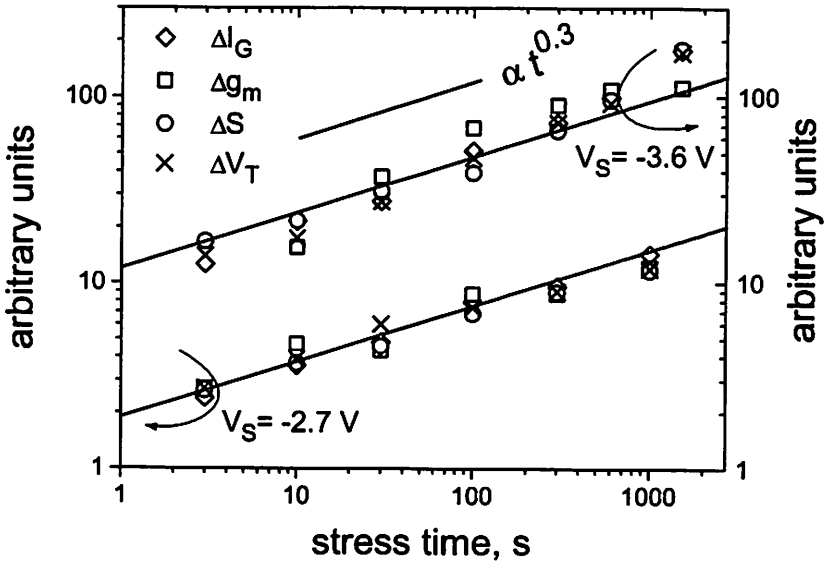
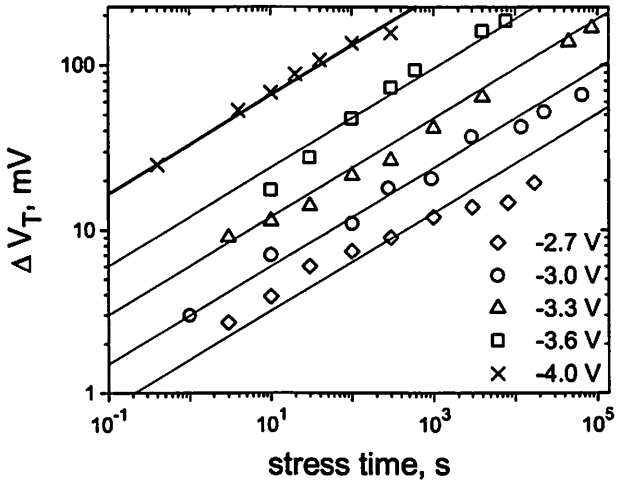


Fig. 2.9. Changes in I_G, g_m, S, V_T all follow the same $t^{0.3}$ dependence on stress time for a wide range of stress voltages.

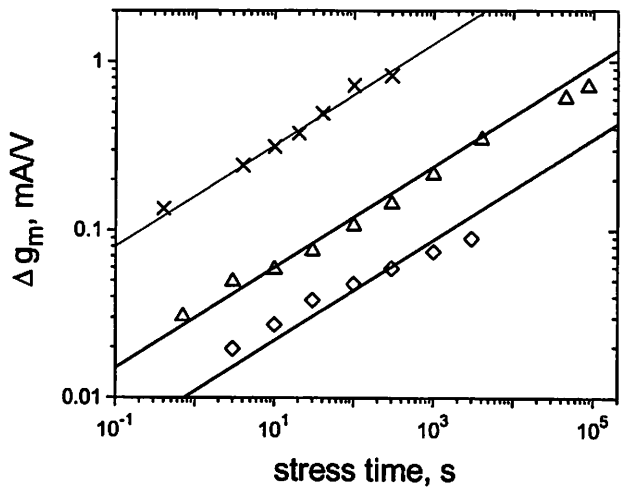
Figure 2.9, the factor $\alpha_Y(V_S)$ depends on the stress voltage, and V_S denotes the absolute value of the stress voltage. The proposed model provides a good fit to the experimental data for ΔI_G, ΔV_T, ΔS, and Δg_m (Fig. 2.10 (a)-(d)). The values of the α_Y are extracted from Figure 2.10 and plotted in Figure 2.11 as functions of the stress voltage. Each of the α_Y's depends exponentially on the stress voltage [2.21, 2.22].

$$\begin{aligned} \Delta V_T &= 3\times10^{-3} e^{2.3V_S} t^{0.3} && mV \\ \Delta g_m / g_m &= 1\times10^{-3} e^{2.0V_S} t^{0.3} && \% \\ \Delta S &= 1\times10^{-4} e^{2.7V_S} t^{0.3} && mV/dec \\ \Delta I_G / I_G &= 2\times10^{-5} e^{3.8V_S} t^{0.3} && \% \end{aligned} \tag{2.2}$$

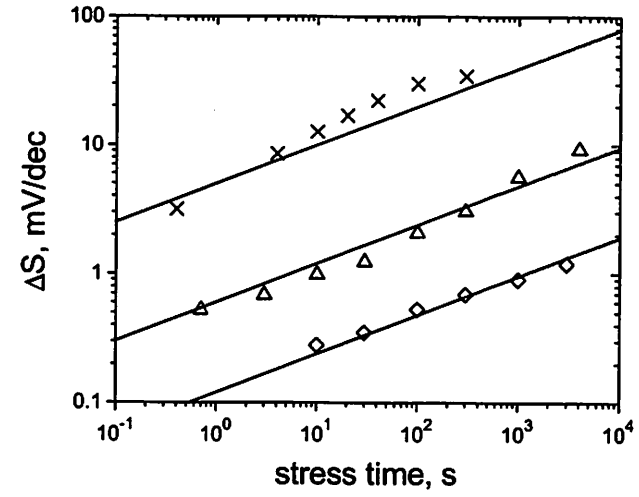
In these expressions, the stress voltage is measured in volts. The changes in the threshold voltage, transconductance, and subthreshold swing all have similar dependence on the stress voltage. This similarity can be attributed to the fact that the build-up of the interface traps is primarily responsible for the deterioration of these three parameters. In contrast, the changes in the gate leakage current have different stress voltage dependence, because a different type of traps (bulk traps in this case) is primarily responsible for the increase in the gate leakage current.



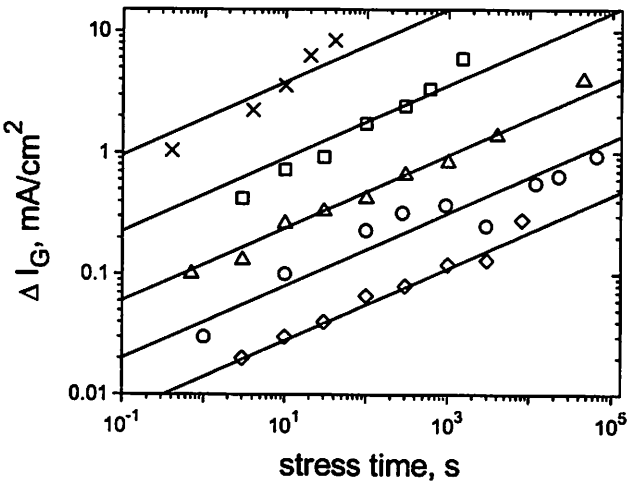
(a)



(b)



(c)



(d)

Fig. 2.10. Equation (2.1) provides good fit for **(a)** ΔV_T, **(b)** Δg_m, **(c)** ΔS, **(d)** ΔI_G as functions of stress time for various stress voltages.

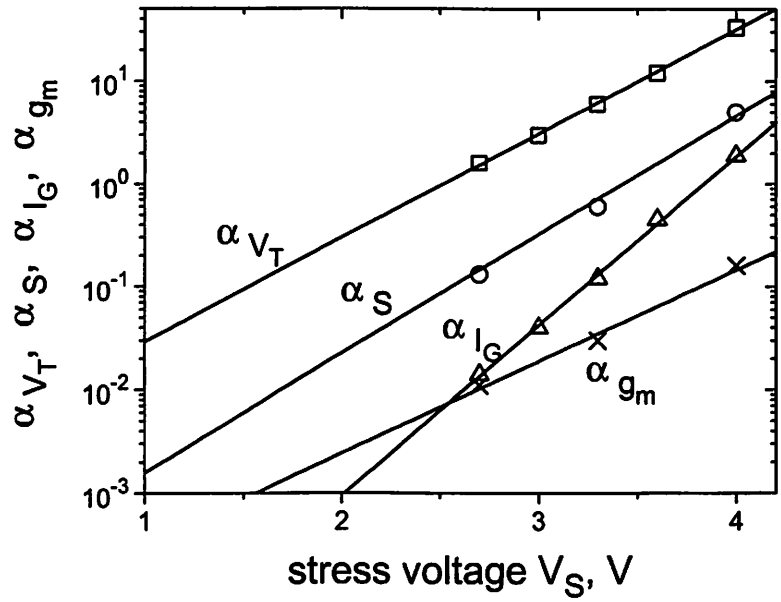


Fig. 2.11. α_Y's vary exponentially with the stress voltage.

The International Technology Roadmap [1.1] predicts that by the year 2005, when the equivalent oxide thickness of the dielectric will scale to below 1.5 nm the circuit supply voltage will be between 0.9 V and 1.2 V depending on the application. Clearly, the higher supply voltage imposes a stricter limit on the reliability. V_S = 1.2V and t = 3×10^8 sec (10 years) were used to project the degradation of the JVD nitride transistors. The results are summarized in Table 2.1. The increase in the gate leakage current becomes negligible at low operating voltages due to the strong voltage dependence of ΔI_G as seen in equation (2.2). The deterioration in other parameters might, to a degree, affect the device performance, but will remain well within the acceptable boundaries.

TABLE 2.1. Projected Changes in the MOSFET Parameters Over a 10-year Device Lifetime under Operating Voltage of 1.2V.

Supply voltage, V_{DD}	1.2 V
ΔV_T	17 mV
Δg_m	4%
ΔI_G	1%
ΔS	1 mV/dec

2.3.5 Physical foundation of the empirical results

The TDDW model for Si_3N_4 described here by the empirical set of equations (2.2) resembles the TDDW model for SiO_2 proposed earlier by Qian and Dumin [2.24], which predicts that the number of traps generated by FN-stress follows an exponential dependence on the applied electric field and a power-law dependence on the stress time. Exponential dependence of trap generation rate on voltage can be explained by either a thermochemical reaction [2.25] or a voltage driven process taking place in a dielectric [2.26, 2.27]. Similar mechanisms are perhaps responsible for trap generation in SiO_2 and Si_3N_4.

The origins of power-law dependence on time are less certain. The hydrogen-atom diffusion model [2.28] has been proposed to explain the power-law dependence with $\beta = 0.5$. However, a range of values of β from 0.2 to 0.75 has been observed. Instead, a simple statistical model could explain the power-law dependence of defect generation. Suppose there is a certain number, N_0, of weak (strained) bonds in the dielectric layer and these bonds are broken under a constant voltage stress with a time constant τ. Then the number of defects generated over time t is $\Delta N = N_0(1\text{-}e^{-t/\tau})$. In this

basic model the number of traps generated is a linear function of time for $t << \tau$ and saturates for $t >> \tau$. In reality however, various traps are likely to have different time constants τ. Although there is currently no clear physical evidence for any particular mechanism responsible for these different time constants, we believe that a number of mechanisms could be responsible. For example, the variations in the trap position within the dielectric and/or the fact that different bonds can be strained to different degrees [2.29] might give rise to different time constants. Interestingly enough, regardless of what this distribution is, the number of traps generated is a power-law (or nearly a power-law) function of time. If one assumes a power-law distribution of τ, for example, (i.e. the number of weak bonds with time constants between τ and $\tau+d\tau$ is $K\tau^n d\tau$), it can be easily shown by integration over all τ's that the number of traps generated will also have a power-law dependence on time.

$$\Delta N = K\int_0^\infty \tau^n\left(1-e^{-t/\tau}\right)d\tau = -\Gamma(-1-n)Kt^{n+1}, \qquad (2.3)$$

where Γ is the gamma-function. In our case $n+1 = 0.3$, and $\Gamma(-0.3)$ is a negative number, so that the right-hand side of equation is positive.

2.4 Hot-carrier stress

2.4.1 Introduction (Background)

The studies of time-dependent dielectric breakdown conducted by other researchers [2.16, 2.30], as well as the study of time-dependent dielectric wearout described earlier in this chapter indicate that Si_3N_4 under Fowler-Nordheim stress meets

reliability requirements. However it still remains to be shown that hot-carrier reliability of Si_3N_4 gate dielectrics is acceptable. Earlier work [2.30] indicates good hot-carrier reliability of NMOSFET JVD nitride transistors with 31 Å equivalent oxide thickness. Next, we will show that the hot-carrier lifetime of Si_3N_4 PMOSFETs is similar to that of SiO_2 PMOSFETs.

We also examine the mechanism responsible for PMOSFET degradation. It has been long known that the mechanism responsible for the device degradation in NMOSFET is interface state generation. The situation for PMOSFETs is less clear. It had long been believed that hot-carrier reliability of PMOSFETs is not as serious an issue as hot-carrier reliability of NMOSFETs for the following reason: The mean free path of holes in silicon is about one half that of the electrons [2.31]; therefore holes scatter more frequently and fewer of them reach high enough energies (about 4 eV) to create interface states [2.32]. However, as the transistor channel length has been scaled down into the deep-sub-micron regime (and supply voltages have been reduced) hot-carrier induced degradation of PMOSFETs has been approaching that of NMOSFETs [2.33]. Consequently, the hot-carrier reliability of PMOSFETs has been studied in more detail. Three hot-carrier degradation mechanisms in PMOSFET's have been identified [2.34, 2.35]. The first is negative oxide charge trapping. Electron trapping near the drain region leads to a reduction in the threshold voltage and to the effective channel shortening. As a result, PMOSFET drive current increases. This mechanism is most important in longer channel PMOSFETs, and gate current I_G has been used as a predictor of the device lifetime. The second mechanism is the generation of interface states by hot holes, which leads to channel mobility degradation. In this case, the substrate current I_{SUB} should be

used to predict the device lifetime. The third mechanism is positive oxide charge trapping. Interface-state generation has been shown to be the dominant degradation mechanism for 0.25μm surface channel PMOSFETs [2.34]. We will show that this conclusion remains true for our 100 nm devices, for both oxide and nitride gate dielectrics.

2.4.2 Hot-carrier measurements

The drain voltage V_D during the hot carrier stress ranged from - 4.5 to - 6.5 V; the gate voltage V_G was chosen to maximize the substrate current. The exact stress conditions for each of the nitride transistors are listed in Table 2.2. As we have outlined in the introduction, hot-electron injection becomes a relatively less important mechanism of PMOSFET degradation as the channel length becomes shorter than 0.25μm [2.34, 2.35]. Therefore, we do not expect the stress at low V_G ($V_G = V_D/5$), which favors hot-electron injection, to be the worst-case stress condition. Many recent papers have been dedicated to the discussion of whether stress at maximum gate voltage ($V_G = V_D$) or stress at maximum substrate current I_{SUB} leads to the fastest device degradation [2.36, 2.37]. We chose to use the stress at maximum I_{SUB} since the stress at maximum V_G would not have been practical in our case. Applying such a high gate voltage to our ultra-thin dielectrics would lead to rapid device degradation under Fowler-Nordheim (rather than hot-carrier) stress. For example, according to the equation (2.2) the transconductance of a JVD transistor under - 4.5 V stress would drop by 10% in about 2 seconds, while the same amount of degradation due to hot carrier stress under peak I_{SUB} condition takes 5×10^5 seconds. (Extrapolated lifetime under - 2 V FN-stress is in excess of 10^7 seconds.) The

hot-carrier stress was periodically interrupted, and transistor parameters such as saturation drain current I_{DSAT}, linear drain current I_{DLIN}, threshold voltage shift ΔV_T, peak transconductance g_m, subthreshold swing S, and gate leakage current I_G were monitored. An example of device degradation is shown in Fig. 2.12. Degradation in transconductance (Fig. 2.12a) and associated degradation in drive current are especially important factors which affect circuit performance. We chose to define the device lifetime as the time to reach 10% degradation in I_{DLIN} (drain current measured at V_G = - 1.5 V and V_D = - 100 mV). I_{DLIN} degradation follows exactly the same time-dependence as the degradation in peak transconductance g_m, while I_{DSAT} changes by 4% for every 10% change in g_m (Fig. 2.13). This is expected since both I_{DLIN} and g_m are directly proportional to hole channel mobility, while I_{DSAT} depends on the hole saturation velocity as well as mobility.

We attribute the increase in the gate leakage current observed in Figure 2.12b to trap-assisted-tunneling. These traps are created in the gate dielectric by hot carriers. This

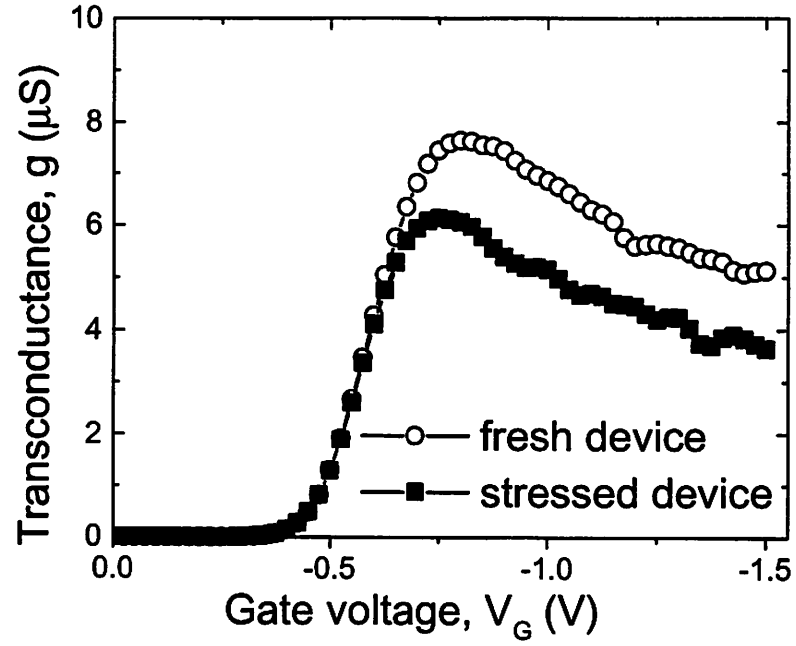
TABLE 2.2. Peak I_{SUB} stress conditions for Si_3N_4 transistors (L/W=0.1/1μm).

Drain voltage V_D (V)	Gate voltage V_G (V)	Peak I_{SUB} (μA)
- 4.5	- 1.2	0.16
- 5.0	- 1.4	0.5
- 5.3	- 1.5	1.0
- 5.5	- 1.6	2.0
- 5.6	- 1.5	2.4
- 6.5	- 2.0	3.6

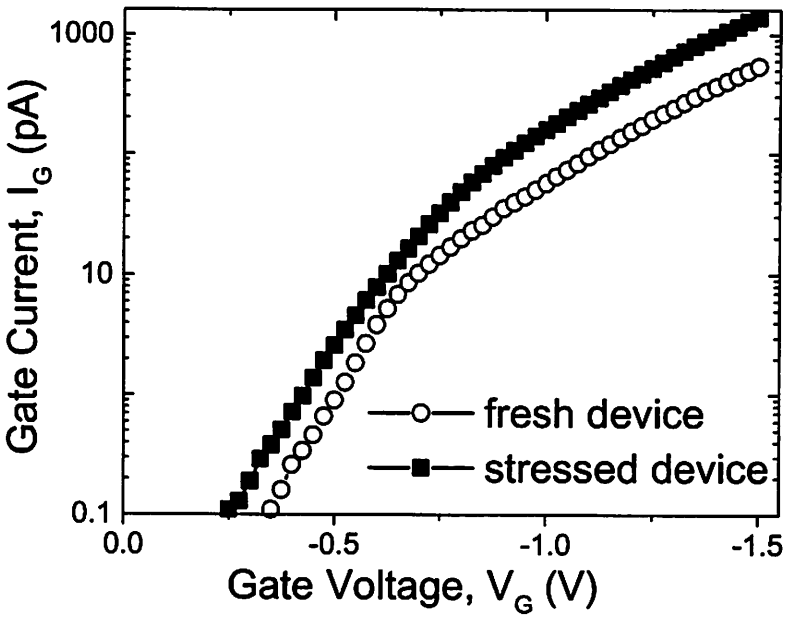
increase in the gate leakage current can become a reliability concern for logic transistors with ultra-thin gate dielectrics, as it can lead to increased power consumption. The changes in V_T will be discussed in the next section.

2.4.3 Degradation mechanism

We examine the mechanism responsible for the device degradation by examining the change in the threshold voltage (Fig. 2.14). In the early stages of stress, electrons are trapped in the gate dielectrics as indicated by the positive ΔV_T. Electron trapping follows a "logarithmic" dependence on time [2.34]. This is consistent with a model in which the electrons are created by impact ionization near the drain region, propelled towards the gate electrode by the vertical electric field, and captured by the traps which exist in the dielectric. Silicon nitride is known to have a higher density of traps than silicon oxide; in addition its physical thickness is almost twice that of the oxide, hence silicon nitride shows a larger positive ΔV_T. In the later stages of stress, ΔV_T becomes negative indicating a positive charge build-up in the gate dielectric. The positive charge can result from either hole trapping in the dielectric or the creation of positively charged interface states at the dielectric interface. In practice, it may be hard to draw a distinction between the two phenomena, as it is hard to distinguish between bulk and interface traps in the case of ultra-thin dielectrics. We believe that interface state generation is predominant, as the electric field near the drain does not favor hole injection in the dielectric. Furthermore, the change in V_T becomes a power-law function of time, consistent with the interface generation model [2.34].



(a)



(b)

Fig. 2.12. Degradation in **(a)** transconductance and **(b)** gate leakage current as a result of hot-carrier stress. (Stress V_D = -6.5V, stress time = 500 sec.)

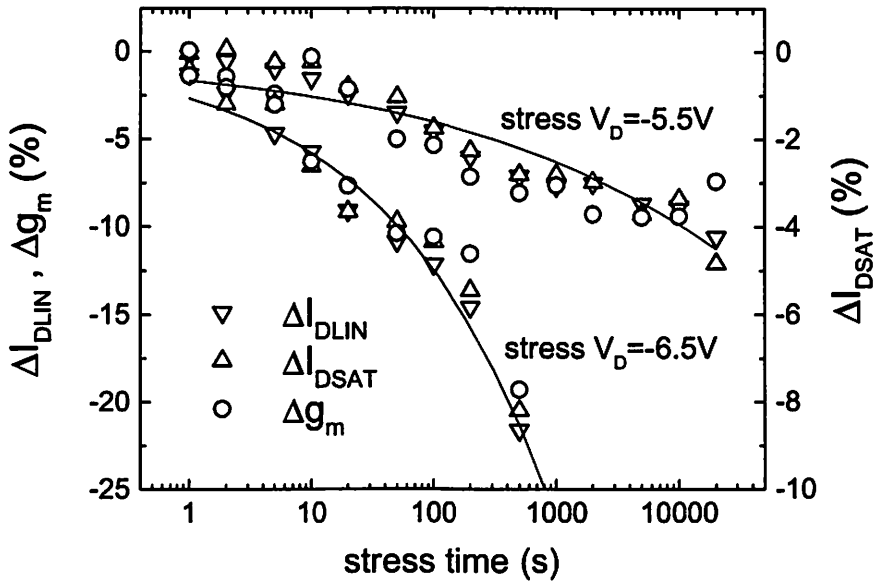


Fig. 2.13. Degradation of Si_3N_4 PMOSFET parameters.

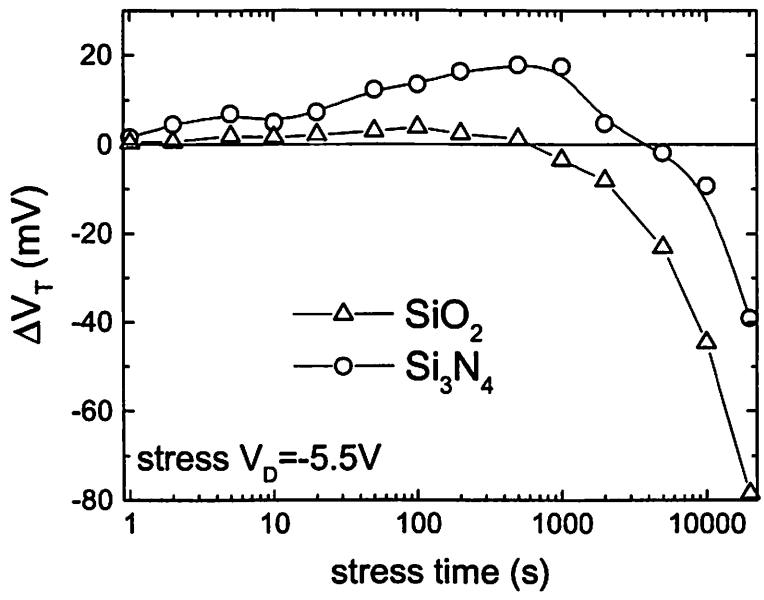


Fig. 2.14. Positive ΔV_T corresponds to electron trapping; negative ΔV_T to hole trapping and interface trap generation.

To further support the interface-state generation model, we note that the result of the charge trapping alone on device performance is quite modest. A 2×10^4 second stress at V_D = - 5.5 V leads to a -40 mV shift in V_T; this would translate into a fraction-of-percent change in I_D. In reality however, drain current changes by more than 10% during this stress. Essentially all of the degradation in I_D is due to the decrease in hole mobility. Interface-state generation is therefore by far the dominant degradation mechanism.

2.4.4 Lifetime comparison for Si_3N_4 and SiO_2 transistors

Our next task is to determine the reliability properties of gate nitride as expressed by the hot-carrier lifetime. In general, hot-carrier lifetime depends on the properties of the gate dielectric as well as on LDD design. Thus simply determining the lifetime of transistors with a new gate dielectric is rather meaningless. Instead, the lifetime of the Si_3N_4 devices should be compared to the lifetime of SiO_2 devices with a similar LDD design. We verified that our nitride and oxide transistors are indeed comparable in terms of source/drain engineering by comparing the substrate currents in these devices. Substrate current is an exponential function of the peak electric field in the channel pinch-off region [2.38]. Since both device types have the same peak I_{SUB} as a function of drain voltage (Fig. 2.15), the electric field in the pinch-off region is the same for these two device types at a given V_D. Therefore we can directly compare the lifetime of the Si_3N_4 transistors against the lifetime of the control SiO_2 devices. Also included in Figure 2.15 is the substrate current measured for SiO_2 devices with a more aggressive (higher doping) LDD design. At a given V_D these devices have a higher peak electric field, and

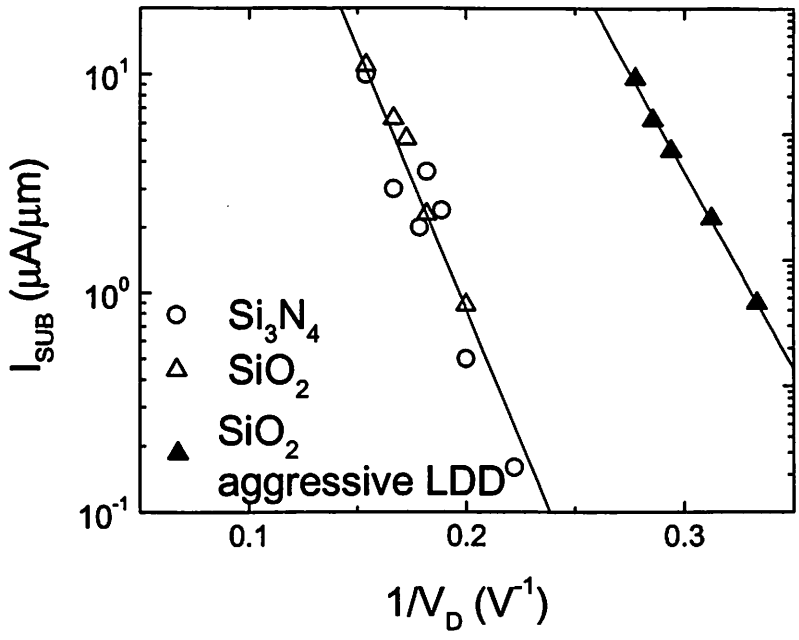


Fig. 2.15. Similar LDD design results in similar I_{SUB} vs. V_D characteristic.

therefore a higher I_{SUB}. Lower lifetime for these devices should be expected at a given V_D. One still should be able to compare the lifetimes of different dielectrics if the comparison is made at the same I_{SUB}. (We have already confirmed that interface state generation by hot holes is the major degradation mechanism; therefore I_{SUB} is the correct lifetime predictor.) Figure 2.16 shows that the lifetime follows the same power-law dependence on I_{SUB} with the slope of 1.5 commonly observed for PMOSFETs [2.32, 2.39] for devices with different gate dielectrics and different LDD designs.

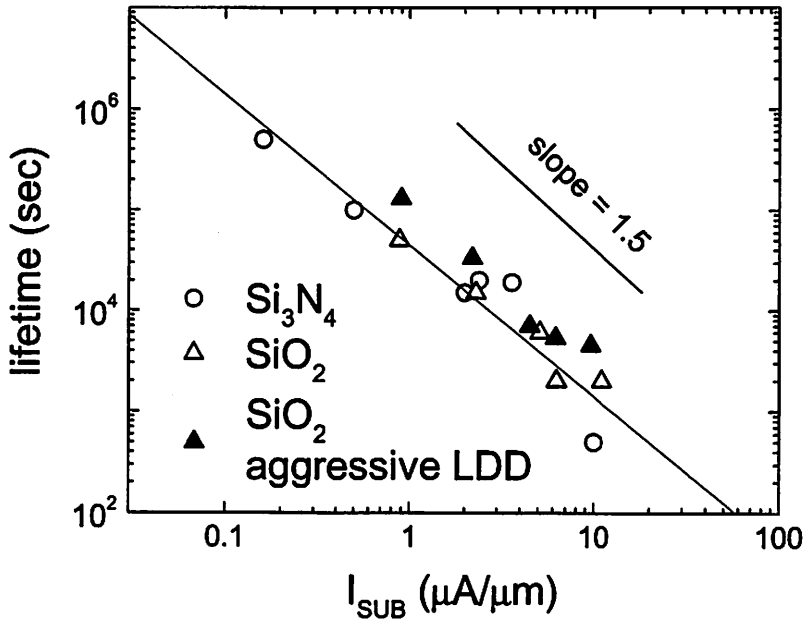


Fig. 2.16. Transistor lifetime depends on I_{SUB}, but not on the gate dielectric or the LDD design.

2.4.5 Lifetime prediction for Si_3N_4 and SiO_2 transistors

We have found that the hot-carrier lifetime of PMOSFETs with JVD nitride gate dielectric is similar to that of devices with SiO_2 gate dielectric. This conclusion is in agreement with the observation by other researchers [2.40] that interface generation in oxide and oxynitride PMOSFETs is insensitive to nitridation. The reason behind this insensitivity is likely to be the fact that hot carriers cause damage to the silicon surface [2.38]. We have also confirmed that interface-state generation remains the dominant device degradation mechanism for deep-sub-micron PMOSFETs. I_{SUB} should therefore be used as the predictor of device lifetime.

The lifetime of transistors as a function of $1/V_D$ is shown in Figure 2.17. The extrapolation to low operating voltages indicates that a supply voltage of around 3.8 V would lead to a 10-year lifetime for both Si_3N_4 and control SiO_2 transistors. This high voltage is explained by the conservative LDD design in our devices. A more aggressive LDD design would lead to a 2.2V supply voltage limit, while improving the transistors' current drive.

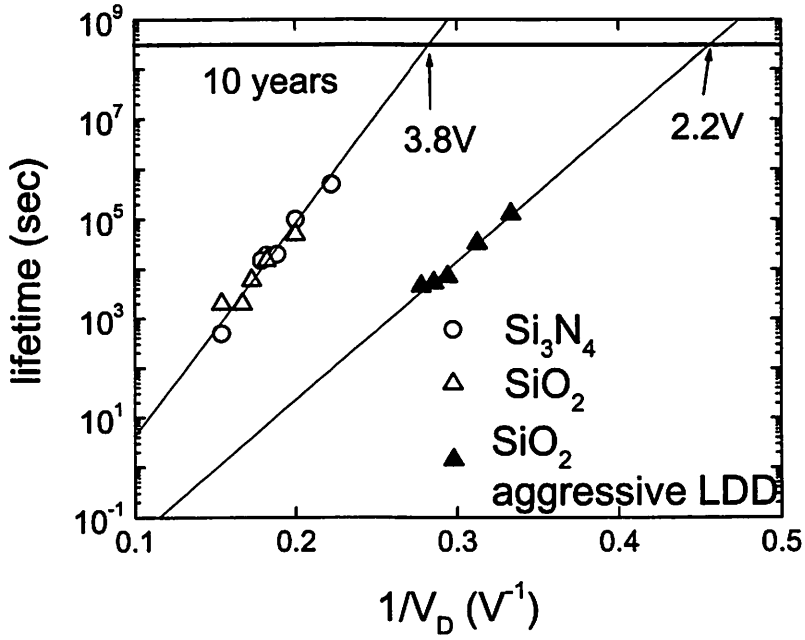


Fig. 2.17. Transistor lifetime extrapolation. L_{eff} = 0.1 μm.

2.5 Random telegraph noise (RTN) model

2.5.1 Background

The switching behavior observed in Figure 2.3, is commonly referred to as random telegraph noise (RTN), because current fluctuates between two levels ("on" and "off" states) and the timing of the switching events is random (typically a random Poisson process). RTN signal have been observed in various MOSFET currents: first in the drain current I_D [2.41, 2.42], and more recently in the gate current I_G [2.43, 2.44].

A good quantitative model of RTN in drain current has been developed. According to the model, the noise originates the oxide-trap-induced fluctuations in the number and mobility of channel carriers. The time spent in the on-state (high-current state) corresponds to the situation when the trap is empty. Therefore, the average time in the on-state is the trap's capture time t_C. Similarly, the time spent in the off-state (low-current state) corresponds to the situation when the trap is occupied, and the average time in the off-state is the trap's emission time t_E. The ratio of capture and emission times can be determined by considering the detailed equilibrium of a trap with respect to the Fermi-level in the substrate. The probabilities of the trap being occupied and empty are $P_O = \frac{1}{1+e^{(E_T-E_F)/kT}}$ and $P_E = 1 - P_O = \frac{e^{(E_T-E_F)/kT}}{1+e^{(E_T-E_F)/kT}}$, where E_T is the trap energy level. The number of transition per unit time from the occupied to empty state $\frac{P_O}{t_E}$ should be equal to the number of transition per unit time from the empty to occupied state $\frac{P_E}{t_C}$, and therefore,

$$\frac{t_C}{t_E} = e^{(E_T - E_F)/kT} \tag{2.4}$$

By changing the gate bias, it is possible to change the position of the trap level with respect to the Fermi level, and therefore change the t_C-to-t_E ratio. By measuring this ratio as a function of the gate voltage, one can determine the position of the trap within the gate dielectric, as shown in Figure 2.18 [2.41].

The RTN in MOSFET's gate current has been observed only recently; and we shall present the first quantitative model describing this phenomenon. We will begin however, with a simple qualitative description of RTN in the gate current of PMOSFET with silicon nitride gate dielectric.

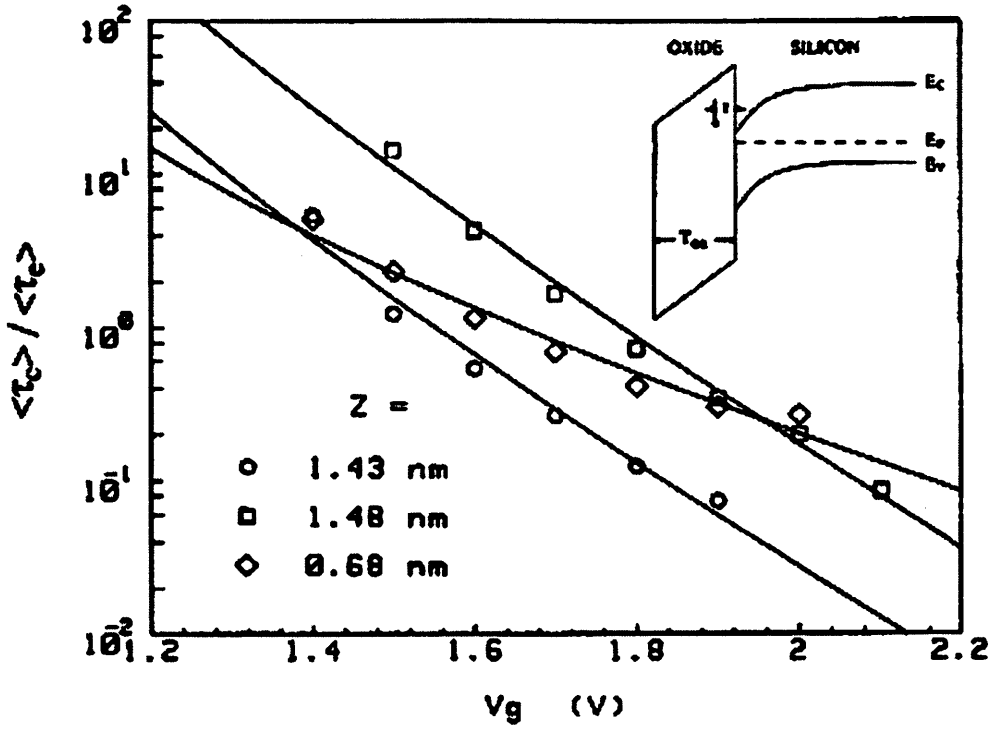


Fig. 2.18. The slopes of the t_C-over-t_E lines allows to determine the position of the traps within 8.6nm oxide [2.41].

2.5.2 Switching behavior: Qualitative description

The PMOSFETs with JVD nitride gate dielectric described earlier in this chapter were studied. After 10^5s of –3V constant-voltage stress, the gate current of a 10μm×10μm transistor increased by about 40% (Fig 2.19). While the amount of stress-induced leakage current (SILC) is modest, post-stress I_G clearly exhibited a telegraph signal characteristic as shown in Figure 2.20. This on-off switching behavior is commonly referred to as noisy or soft breakdown. We propose the following model to explain how SILC and noisy breakdown are linked.

A single bulk trap generated by stress is substantially responsible for the switching behavior observed in Figure 2.20. Let us first consider the off-state (Fig. 2.21a), that is when the trap is empty and therefore carries a single positive charge. The potential of a charged trap in Si_3N_4 has a number of higher states (excited states) in addition to the deep ground state. It is commonly agreed that SILC is the result of trap-assisted tunneling (TAT) through bulk traps [2.45]. The electron tunneling time from an excited state can be estimated to be in the nanosecond range when the trap is located in the middle of the gate dielectric. This is consistent with the observed steps in the current level of 0.3 nA (or 1 electron per 0.5ns). The electron capture (and emission) times for the ground state are much longer since they rely on unlikely multi-phonon processes.

Now let's consider the case when an electron is finally captured in the ground state; and the occupied trap becomes neutral (Fig. 2.21b). The potential barrier is now raised substantially; and this prevents other electrons from tunneling. As a result, trap-assisted tunneling is blocked, and I_G drops. The trap remains occupied on average for a time t_E, then the electron is finally emitted, TAT resumes, and I_G rises again.

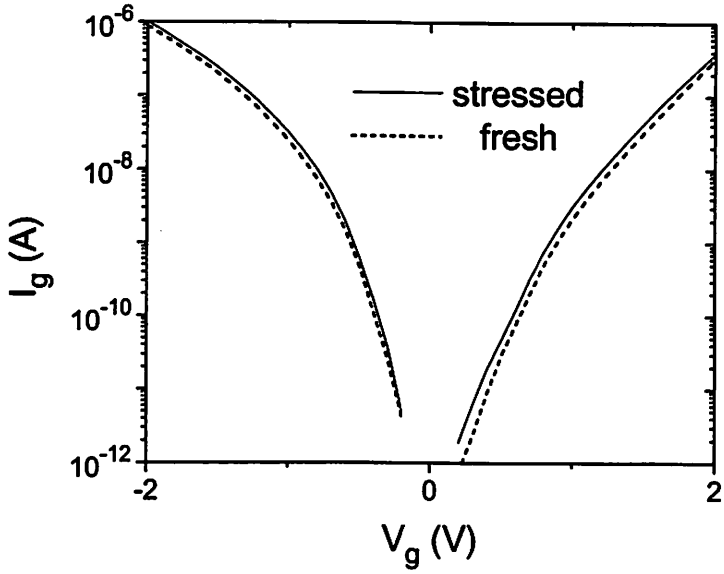


Fig. 2.19. Stress-induced leakage current (SILC) in 1.4 nm silicon nitride film.

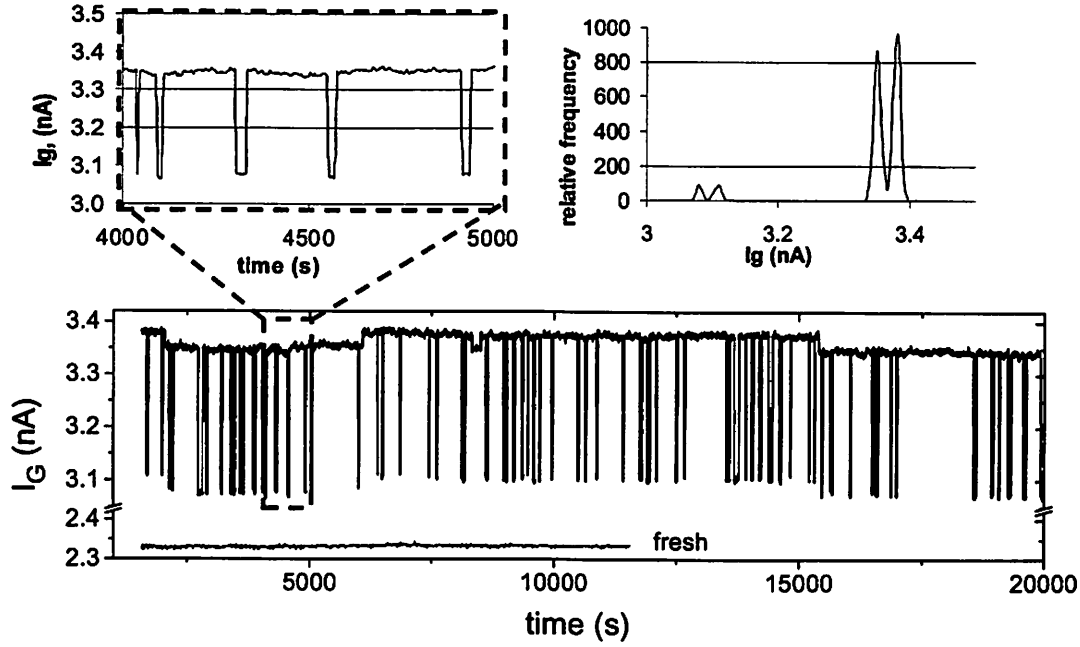


Fig 2.20. Random telegraph noise observed at V_G =1 V. Fig. 2.20b (in the upper right corner) shows a histogram of I_G distribution. 4 peaks correspond to the 4 distinct current levels that are attributed to 2 independent switching traps.

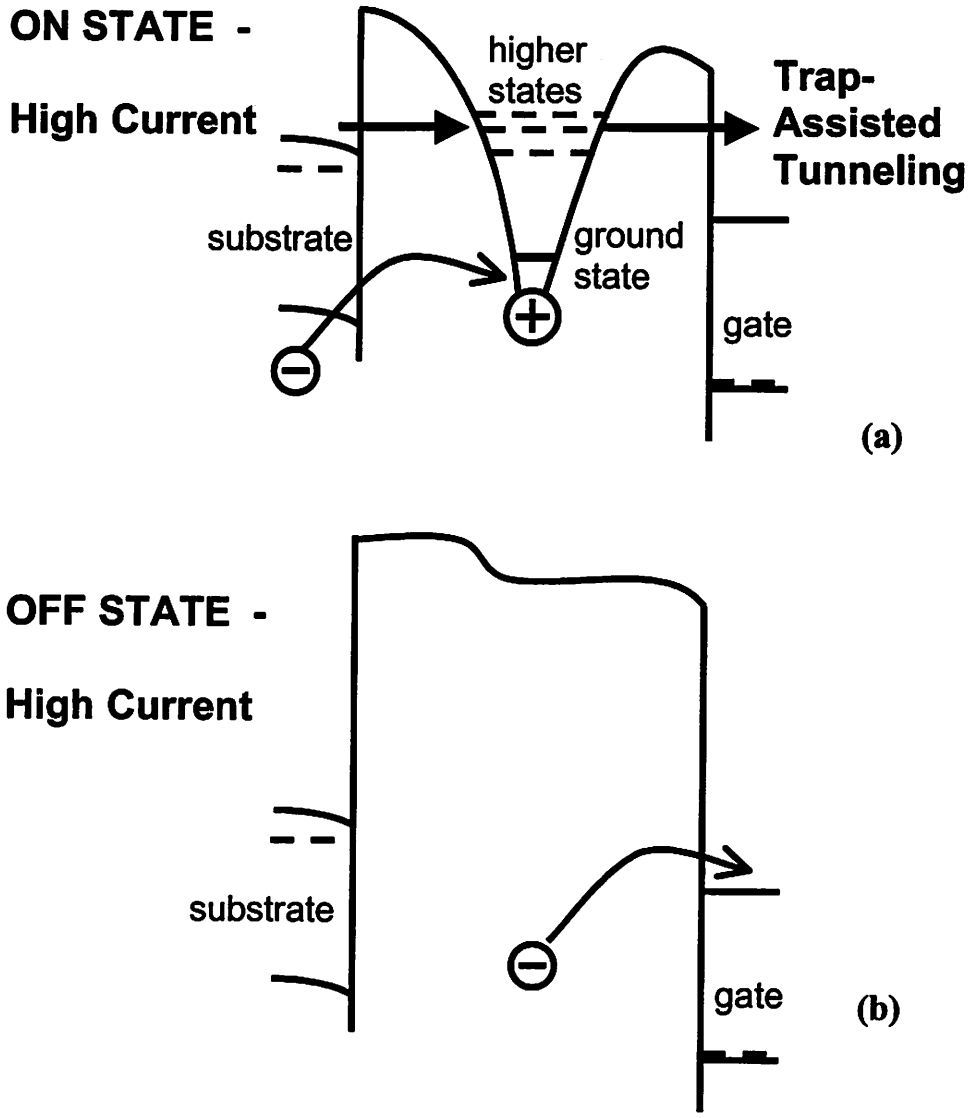


Fig. 2.21. **(a)** Gate dielectric with an empty positively charged trap allows for trap-assisted tunneling current. When a valence band electron is captured from the substrate the trap becomes neutral **(b)** and the trap-assisted tunneling stopps. Eventually the electron is emitted to the conduction band in the gate, and that completes one switching cycle.

The trap that we have just described is located near the middle of the gate dielectric (as we shall show later), and therefore acts as a strong conduction path. A second trap (conduction path) contributing 0.03nA steps can also be observed in Figure 2.20. Many even weaker paths are contributing to the low amplitude noise. The link between SILC and the noisy breakdown (or RTN) now becomes clear: A large number of weak conduction paths (off-center traps) contribute to SILC. Noisy breakdown occurs when a strong tunnel path is created by electrical stress.

Only large conduction paths can be distinguished. There are two such paths in this sample: one contributing 0.3 nA steps, the other – 0.03 nA steps. This explains the doublets in Figure 2.20b (4 current levels correspond to 4 possible states of the two traps: on-on, on-off, off-on, and off-off). After additional stress, additional weak conduction paths are created and the total level of leakage current increases while the contribution from the two larger paths remains unchanged (Fig. 2.22). One can also notice the broadening of the peaks. This is due to the fact that new weak conduction paths increase the amount of the low-amplitude noise in the gate current.

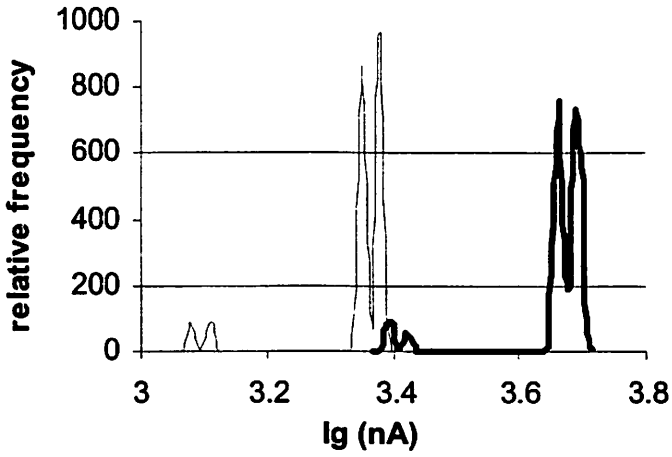


Fig 2.22. As additional weak conduction paths are created, the histogram peaks are shifted to the right and broadened.

2.5.3 RTN bias dependence – Trap location

We have described in section 2.5.1, how the ratio of the capture and emission times for a dielectric trap change with the applied bias. According to (2.4), that ratio changes by one order of magnitude for each 60 mV change in the position of the trap level with respect to the Fermi level in the substrate (60 mV = $kT\ ln10$). In our case however, the ratio of the capture and emission times remains constant (Fig 2.23). There is an important difference that accounts for this discrepancy with the previously developed model. Since the surface traps, described in 2.5.1, exist in the vicinity of the substrate, their occupation probability is determined by the substrate Fermi level. In case of the trap-assisted tunneling, we deal with a trap located in the bulk of the gate dielectric. Since the trap captures electrons from the substrate (Fig. 2.21a), it is the substrate Fermi level that determines the capture time. The emission time, on the other hand, is determined by the Fermi level in the gate, because trapped electrons are emitted into the gate (Fig. 2.21b).

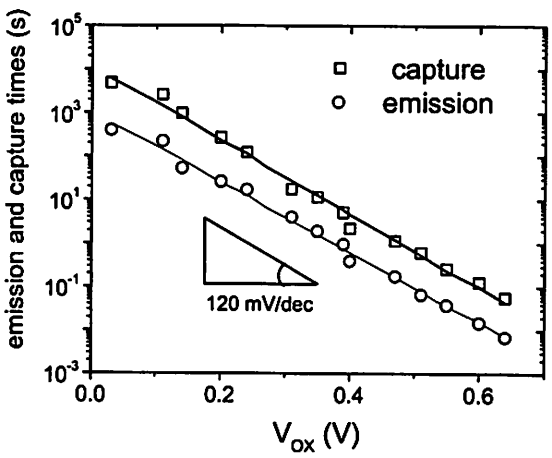


Fig. 2.23. The capture and emission times (t_C and t_E) both change by a factor of 10 for every 120 mV change in the oxide voltage. This means that the trap is located in the middle of the gate dielectric.

If we assume that our bulk trap is located precisely in the middle of the gate dielectric, then a 120 mV increase in V_{OX} will move the trap energy level 60 mV further away from the Fermi levels in both the gate and the substrate. Consequently, we would expect both emission and capture times to increase by a factor of 10 for every 120 mV increase in V_{OX}. This is precisely what we observe in Figure 2.23. Hence we conclude that this trap is indeed located in the middle of the gate dielectric.

2.5.4 RTN temperature dependence – Trap energy

To explore the nature of the bulk traps further we investigated the temperature dependence of RTN. Figure 2.24 shows I_G versus time measured at different temperatures. The amount of current flowing through a single trap (~1nA) is rather insensitive to temperature change as expected for a tunneling current (Fig. 2.25). The capture and emission rates on the other hand increase exponentially with temperature (Fig. 2.26). This strong temperature dependence is expected since high-energy phonons are required to facilitate the capture and emission processes. The activation energies for the capture and emission processes can be extracted from Figure 2.26.

According to Boltzmann statistics, the activation energy for a multi-phonon process E_A is equail to nE_{PH}, where n is the number of phonons involved and E_{PH} is the energy of a single phonon. Typical phonon energy of about 50-60 meV is comparable to kT, and consequently it is important to consider the results of Bose-Einstien statistics, which in reality governs the distribution of phonons:

$$E_A = nE_{PH}\left(1+\frac{1}{e^{E_{PH}/kT}-1}\right) \approx 1.2nE_{PH} \qquad (2.5)$$

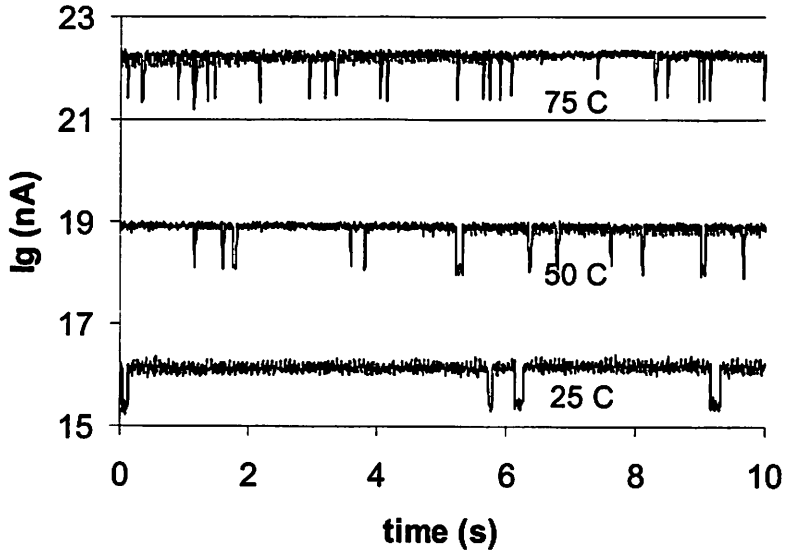


Fig. 2.24. Capture and emission rates increase with temperature. (V_G=1.3V, V_{OX}=0.4V).

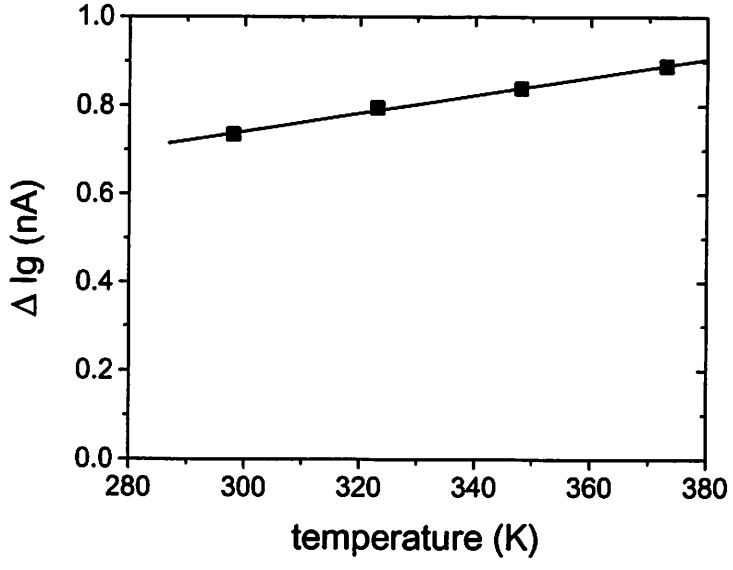


Fig. 2.25. Current tunneling through a single trap is relatively insensitive to temperature. (V_G=1.3V, V_{OX}=0.4V).

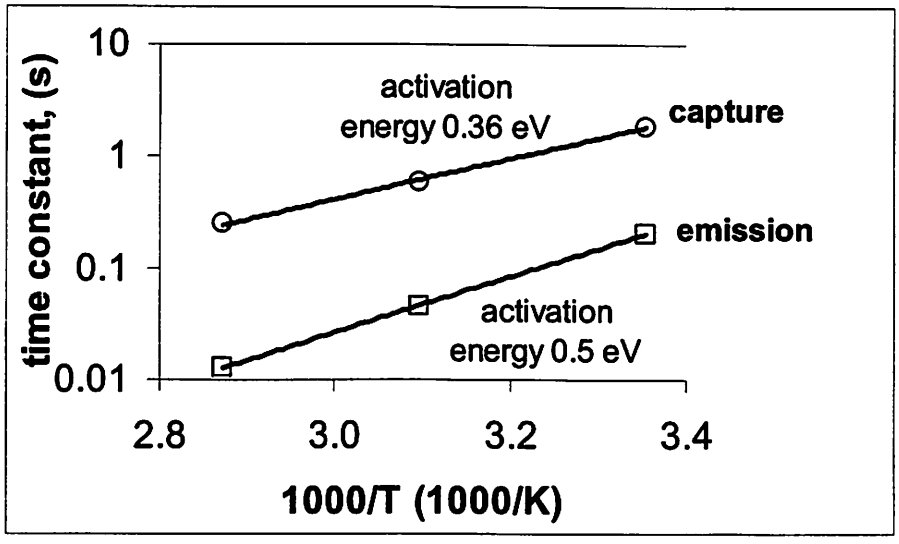


Fig. 2.26. Capture and emission times depend exponentially on temperature. Activation energies roughly correspond to the energies absorbed by an electron during capture and emission. (V_G=1.3V, V_{OX}=0.4V).

This means that the measured activation energy is 20% higher than total energy of the phonons required for emission (capture) processes. Thus the activation energy of 360 meV for the capture process indicates that the trap level is located 300 meV above the substrate's valence band (Fig 2.27). Similarly, the activation energy of 500 meV for the emission process indicates that the trap level is located 420 meV below the gate's conduction band. It can be confirmed that the model is self-consistent, because when we add 300 meV, 420 meV, and the oxide voltage of 400 meV, we get precisely the Si bandgap of 1.12 eV (Fig 2.27). Consequently, we can determine the position of the trap level with respect to the conduction band of Si_3N_4. The trap level is located 2.75 eV below the Si_3N_4 conduction band. A trap level at that energy in Si_3N_4 has been previously reported [2.46].

Finally, we note that the observed traps can be easily annealed out at a slightly elevated temperature. Figure 2.28 show gate current measured at 100°C, for the same trap that is shown in Figures 2.20 and 2.24. The random telegraph switching occurs up to time t=72 seconds. Thereafter, the trap-assisted tunneling no longer occurs, because the trap has been annealed out.

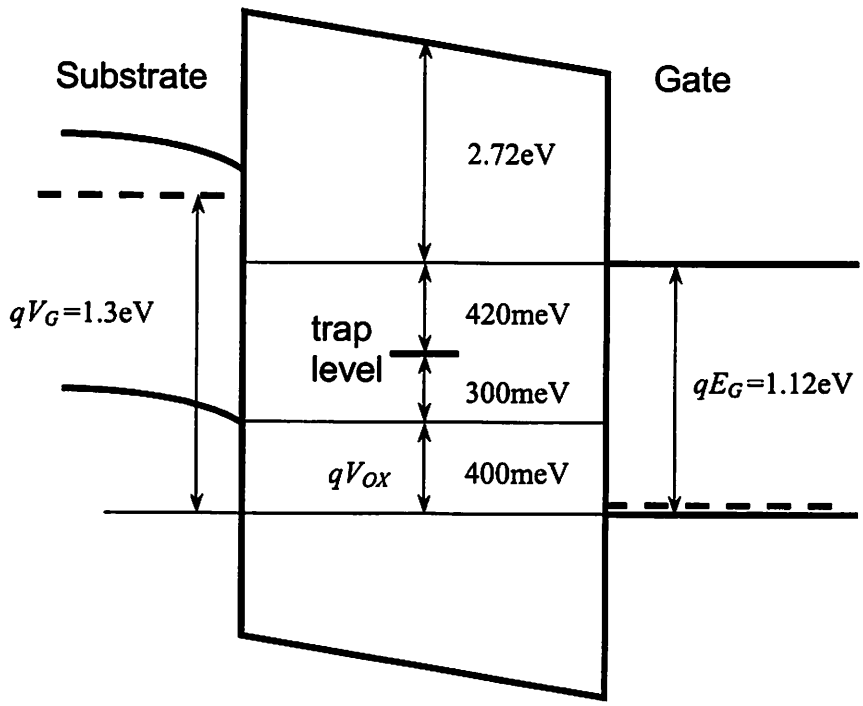


Fig. 2.27. Band diagram of the gate dielectric biased at V_G=1.3V (V_{OX}=0.4V). The energy the trap level with respect to the gate, substrate, and the conduction band edge of the dielectric is indicated. (Dielectric barrier heights are not drawn to scale.)

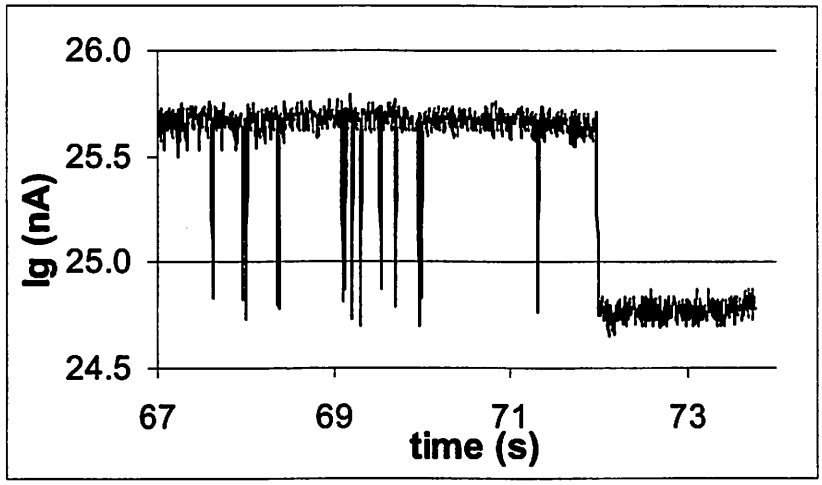


Fig. 2.28. The large conductive path disappears, as the trap is annealed out at 100 °C (V_G=1.3V).

2.5.5 Summary

In summary, a large number of weak conduction paths (off-center traps) contribute to SILC in silicon nitride. Soft or noisy breakdown occurs when a strong tunnel path (in our 1.4 nm nitride, a single trap located near the center of the film) is created by electrical stress. The ground state of the trap is 2.75eV below the E_C of the nitride. When the ground state is empty, electrons from the cathode tunnel to the anode assisted by the higher energy states at the rate of one electron per 0.5ns. When the ground state captures an electron from the cathode, TAT is turned off until the captured electron tunnels out into the anode. The proposed model explains the physical origin of SILC and noisy breakdown, and can be applied to Si_3N_4, SiO_2 and other gate dielectrics.

2.6 Conclusion

The intrinsic reliability of a 1.4 nm (equivalent oxide thickness) JVD gate nitride has been investigated. It has been confirmed that thin JVD nitride is not susceptible to hard dielectric breakdown at low operating voltages. At the same time a drift in MOSFET performance due to the FN-stress has been observed. A model that describes the degradation of the MOSFET parameters as a function of time and stress voltage has been proposed. This model allows the prediction of the device degradation operating under a low supply voltage over a long period of time. According to these projections the degradation of the device performance remains well within the allowed range.

Creation of the interface states at the Si surface has been identified as a leading cause of the PMOS hot-carrier degradation. Because the damage is likely caused to the Si surface itself, we observe that hot-carrier reliability of Si_3N_4 transistors is similar to that of SiO_2 transistors. In addition, a 14Å Si_3N_4 PMOSFET has higher I_{DSAT} and lower gate leakage current than a 16Å SiO_2 PMOSFET.

The results of these reliability studies indicate that Si_3N_4 is a promising alternative to SiO_2 as the dielectric of choice for future generations of CMOS devices.

2.7 References

2.1 Robin Degraeve, Ben Kaczer, Guido Groeseneken, "Degradation and breakdown in thin oxide layers: mechanisms, models and reliability prediction," *Microelectronics Reliability,* Vol. 39, pp. 1445-1460, Oct. 1999.

2.2 Chun-Yen Chang, Chi-Chun Chen, Horng-Chih Lin, Mong-Song Liang, Chao-Hsin Chien, Tiao-Yuan Huang, "Reliability of ultrathin gate oxides for ULSI devices," *Microelectronics Reliability,* Vol. 39, pp. 553-566, 1999.

2.3 J. Sune, E. Y. Wu, D. Jimenez, R. P. Vollertsen, E. Miranda, "Understanding soft and hard breakdown statistics, prevailing ratios and energy dissipation during breakdown runaway," *in IEDM Tech. Digest*, pp. 117-120, 2001

2.4 T. Tsuchiya. Y. Okazaki, M. Miyake, T. Kobayashi, "New hot-carrier degradation mode and lifetime prediction method in quarter-micrometer PMOSFET," *IEEE Trans. on Electron Devices*, Vol. 39, p. 404, 1992.

2.5 R. Woltjer, G. M. Paulzen, H.G. Pomp, H. Lifka, P.H. Woerlee, "Three hot-carrier degradation mechanisms in deep-submicron PMOSFET's," *IEEE Trans. on Electron Devices*, Vol. 42, p. 109, 1995.

2.6 The International Technology Roadmap for Semiconductors, Semiconductor Industry Association, San Jose, CA, 2001.

2.7 J. H. Stathis and D. J. DiMaria, "Reliability projection for ultra-thin oxides at low voltages," *in IEDM Tech. Dig.,* pp. 167-710 1998.

2.8 Robin Degraeve, Ben Kaczer, Guido Groeseneken, "Reliability: a possible showstopper for oxide thickness scaling?" *Semicond. Sci. Technol.,* Vol. 15, pp. 436-444, 2000.

2.9 B. H. Lee et al., "Ultrathin hafnium oxide with low leakage and excellent reliability for alternative gate dielectric application," *in IEDM Tech. Dig.*, pp. 133-136, 1999.

2.10 H. F. Luan et al., "High quality Ta_2O_5 gate dielectric with $T_{ox,eg} < 10$ A," *in IEDM Tech. Dig.*, pp. 141-144, 1999.

2.11 W.-J. Qi *et al.*, "MOSCAP and MOSFET characteristics using ZrO_2 gate dielectric deposited directly on Si," *in IEDM Tech. Dig.*, p. 145-148, 1999.

2.12 T. P. Ma, "Making silicon nitride film a viable gate dielectric," *IEEE Trans. on Electron Devices*, Vol. 45, p. 680, 1998.

2.13 L. C. Parrillo, J. R. Pfiester, J.-H. Lin, E. O. Travis, and R.D. Sivan, "An advanced 0.5 μm CMOS disposable LDD spacer technology," *in Symp. VLSI Tech. Dig.*, p. 31, 1989.

2.14 available at http://www-device.eecs.berkeley.edu/qmcv.html

2.15 Y.-C. Yeo, Q. Lu, W.-C. Lee, T.-J. King. C. Hu, X. Wang, X. Guo, T.P. Ma, "Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric," *IEEE Electron Device Lett.*, Vol. 21, p.540, 2000.

2.16 Q. Lu *et al.*, "Comparison of 14Å $T_{OX,EQ}$ JVD and RTCVD silicon nitride gate dielectrics for sub-100nm MOSFETs," *Int. Semiconductor Device Research Symp.*, pp. 489-492, 1999.

2.17 M. Khare, X. W. Wang, and T. P. Ma, "Highly robust ultra-thin gate dielectric for giga scale technology," *Symp. VLSI Tech. Dig. Papers*, p. 218-9, 1998.

2.18 K. Sekine *et al.*, "Highly-integrity ultra-thin silicon nitride film grown at low temperature for extending scaling limit of gate dielectric," *in IEDM Tech Dig.*, pp. 115-118, 1999.

2.19 Y. Wu, X. Qi, D. Bang, G. Lucovsky, M.-R. Lin, "Time dependent dielectric wearout (TDDW) technique for reliability of ultrathin gate oxides," *IEEE Trans. Electron Devices*, Vol. 20, pp. 262-264, June 1999.

2.20 C.-Y. Chang *et al.*, "Reliability of ultrathin gate oxides for ULSI devices," *Microelectronics Reliability*, Vol. 39, pp. 553-556, 1999.

2.21 C. Hu *et al.*, "Hot-electron-induced MOSFET degradation – model monitor, and improvement," *IEEE Electron Device Lett.*, Vol. ED-32, pp. 375-85, 1985.

2.22 K. Okada, H. Kubo, A. Ishinaga, K. Yoneda, "A new prediction method for oxide lifetime and its application to study dielectric breakdown," *Symp. VLSI Tech. Dig. Papers*, pp. 158-159, 1998.

2.23 M. Khare, X. W. Wang, and T. P. Ma, "Highly robust ultra-thin gate dielectric for giga scale technology," *Symp. VLSI Tech. Dig. Papers*, pp. 218-220, 1998.

2.24 D. Qian, D.J. Dumin, "A comprehensive physical model of oxide wearout and breakdown involving trap generation, charging, and discharging," *IEEE International Integrated Reliability Workshop Final Report*, pp. 55-61, Lake Tahoe, CA, 12-15 Oct. 1998.

2.25 J. W. McPherson, V. K. Reddy, and H. C. Mogul, "Field-enhanced Si–Si bond-breakage mechanism for time-dependent dielectric breakdown in thin-film SiO2 dielectrics" *Applied Physics Letters*, Vol. 71, pp. 1101-1103, 1997.

2.26 P.E. Nicollian, W. R Hunter, J. C. Hu, "Experimental evidence for voltage driven breakdown models in ultra thin gate oxides," *Int. Reliability Phys. Symp.*, p. 7, 2000.

2.27 M. A. Alam, J. Bude, A. Ghetti, "Field acceleration for oxide breakdown – Can an accuate anode hole injection model resolve the E and 1/E controvercy?" *Int. Reliability Phys. Symp.*, p. 21, 2000

2.28 D. Arnold, E. Cartier, D.J. DiMaria, "Theory of high-field electron transport and impact ionization in silicon dioxide," *Physical Review B (Condensed Matter)*, Vol.49, No.15, pp. 10278-10297, 15 April 1994.

2.29 G. Bersuker, Yongjoo Jeon, H. R. Huff, "Degradation of thin oxides during electrical stress," *Microelectronics Reliability*, Vol. 41, No. 12, pp. 1923-31, Dec. 2001.

2.30 S. Mahapara *et al.*, "100 nm channel length MNSFETs using a jet vapor deposited ultra-thin silicon nitride gate dielectric," *Digest of Technical Papers, 1999 Symp. on VLSI Technology*, p79.

2.31 T.-C. Ong, P.-K. Ko, C. Hu, "Modeling of substrate current in p-MOSFET's," IEEE Electron Device Lett., EDL-8, p. 413, 1987.

2.32 T.-C. Ong, P.-K. Ko, C. Hu, "Hot-carrier current modeling and device degradation in surface-channel p-MOSFET's," *IEEE Trans. on Electron Devices*, Vol. 37, p. 1658, 1990.

2.33 T. Tsuchiya. Y. Okazaki, M. Miyake, T. Kobayashi, "New hot-carrier degradation mode and lifetime prediction method in quarter-micrometer PMOSFET," *IEEE Trans. on Electron Devices*, Vol. 39, p. 404, 1992.

2.34 R. Woltjer, G. M. Paulzen, H.G. Pomp, H. Lifka, P.H. Woerlee, "Three hot-carrier degradation mechanisms in deep-submicron PMOSFET's," *IEEE Trans. on Electron Devices*, Vol. 42, p. 109, 1995.

2.35 A. Bravaix, "Hot-carrier degradation evolution in deep-submicrometer CMOS technologies," *IEEE International Integrated Reliability Workshop, Final Report*, pp. 174-183, 1999.

2.36 A. Bravaix, D. Vuillaume, D. Goguenheim, V. Lasserre, M. Haond, "Competing AC hot-carrier degradation mechanisms in surface-channel p-MOSFET's during pass transistor operation," *in Proccedings of International Electron Device Meeting*, p. 873, 1996.

2.37 E. Li, E. Rosenbaum, L. F. Register, J. Tao, P. Fang, "Hot carrier induced degradation in deep submicron MOSFETs at 100°C," *International Reliability Phys. Symp.*, p. 103, 2000.

2.38 C. Hu, S.C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, K.W. Terrill, "Hot-electron-induced MOSFET degradation – model, monitor, and improvement," *IEEE Trans. on Electron Devices*, ED-32, p. 375, 1985.

2.39 C.H. Liu, M.G. Chen, S. Huang-Lu, Y.J Chang, K.Y. Fu, "Analysis of hot-carrier degradation in 0.25 μm surface-channel pMOSFET device," *Digest of Technical Papers, Symp. on VLSI Technology*, p. 82, 1999.

2.40 J.F. Zhang, H.K. Sii, G. Groeseneken, R. Degraeve, "Degradation of oxides and oxynitrides under hot hole stress," *IEEE Trans. on Electron Devices*, Vol. 47, p. 378, 2000.

2.41 K.K. Hung, P.K. Ko, C. Hu, Y.C. Cheng, "Random telegraph noise of deep-submicrometer MOSFETs," *IEEE Electron Device Letters*, Vol. 11, No.2, pp. 90-92, Feb. 1990.

2.42 K.K. Hung, P.K. Ko, C. Hu, Y.C. Cheng, "A unified model for the flicker noise in metal-oxide-semiconductor field-effect transistors," *IEEE Transactions on Electron Devices*, Vol. 37, No.3, Pt.1, pp. 654-665, March 1990.

2.43 F. Crupi, R. Degraeve, G. Groeseneken, T. Nigam, H.E. Maes, "On the properties of the gate and substrate current after soft breakdown in ultrathin oxide layers," *IEEE Transactions on Electron Devices*, Vol. 45, No.11, pp. 2329-2334, Nov. 1998.

2.44 O. Briere, J. A. Chroboczek, G. Ghibaudo, (Edited by: G. Baccarani, M. Rudan) "Random telegraph signal in the quasi-breakdown current of MOS capacitors," in *Proceedings of the 26th European Solid State Device Research Conference*, Bologna, Italy, Gif-sur-Yvette, France: Editions Frontieres, pp. 759-762. 1996.

2.45 E. Rosenbaum, L. F. Register, "Mechanism of stress-induced leakage current in MOS capacitors," *IEEE Transactions on Electron Devices*, Vol. 44, No. 2, pp. 317-323, Feb 1997.

2.46 V. J. Kapoor, The physics of MOS insulators, Pergamon Press, New York, p. 117, 1980.

Chapter 3 Tunneling Model for Multi-Layer Gate Dielectrics

3.1 Introduction

Further aggressive reduction in the electrical thickness of the MOS gate dielectrics is required to control the short channel effects, especially for bulk MOSFETs. The major challenge to continued scaling of the gate SiO_2 is posed by the exponentially increasing gate leakage current. The increasing leakage current results in the increased stand-by power consumption which, in turn, leads to increased power dissipation on the chip and shorter battery life for portable devices. Since an understanding of leakage current is so critical for future CMOS circuits, a simple and precise model that predicts leakage current through the gate dielectrics is an indispensable tool in analyzing the scalability of MOS devices.

In order to suppress the gate tunneling current, materials with dielectric constants higher than that of SiO_2 have to be introduced into the gate stack. Silicon nitride and oxynitride films have already become a common part of the gate stack for high-performance MOS devices. Stricter leakage limits imposed on the low-power circuits will eventually require the introduction of true high-permittivity dielectrics, such as HfO_2.

Precise analytical models have been developed to model tunneling through SiO_2 [3.1] as well as other gate dielectrics [3.2]. These semi-empirical models have made possible the evaluation of scaling limits for various single-layer gate dielectrics.

It is likely however, that in order to ensure high channel mobility a thin interfacial layer of SiO_2 or oxynitride will have to be used in conjunction with high-κ dielectric. It therefore becomes evident that the future gate dielectric stack will consist of at least two layers. The task of calculating the tunneling current through the multi-layer dielectric stacks has been addressed by both numerical and analytical means [3.3-3.6]. However the analytical expressions, that are usually obtained by WKB approximation, are rather complex so that the results are typically presented by plotting gate current versus gate voltage for a given gate stack thickness and composition. The results presented in such form are hard to use to project the scalability of the gate stacks, since the leakage current would have to be recalculated for each dielectric thickness. In this chapter, we shall develop a model that result in a compact analytical expression for gate leakage current as a function of the equivalent oxide thickness. Such a model is a much more convenient tool for predicting the scaling limits for multi-layer gate dielectric stacks.

3.2 Modeling approach

While the tunneling of inversion layer electrons through the gate dielectric is clearly a quantum mechanical phenomenon, it is convenient to address the task of modeling the tunneling current in a semi-classical way. The electrons that are confined to the inversion layer potential well can be thought of as oscillating classical particles. (The precise oscillation frequency, of course, has to be evaluated based on the true quantum-

mechanical distribution of the electron wave function.) We shall refer to the oscillation frequency as the impingement frequency, f_{IMP}, since this value represents how often electrons impinge at the dielectric interface. The flux of the impinging electrons is then the product of f_{IMP} and the inversion charge density, Q_{INV}. The gate tunneling current density, J_G, is in turn the product of this flux and the tunneling probability, T:

$$J_G = Q_{INV} f_{IMP} T \tag{3.1}$$

Q_{INV} is a simple function of the gate and threshold voltages: $Q_{INV} = C_{OX}(V_G - V_T)$. As we shall show in the next section, f_{IMP} is also a function of V_G and V_T.

It should be noted that in reality two types of electrons, heavy and light, are confined in the inversion layer, and therefore it is, strictly speaking, necessary to determine Q_{INV}, f_{IMP} and T for each type of the electrons separately before addeding up the two resulting leakage currents. In this case, however, the Q_{INV}'s and f_{IMP}'s will no longer be simple functions of V_G and V_T, making it difficult to come up with a simple analytical model for leakage current. Fortunately, since the energy levels of the heavy electrons are lower, most of the electrons in the inversion layer are heavy electrons. This is especially true at larger gate biases, where the modeling of the leakage current is most important. We shall hereafter assume that only one type of electrons with effective mass m_l=0.96 is present in the inversion layer.

3.3 Analysis of inversion layer charge confinement

3.3.1 Effective electric field

A typical shape of the potential well that confines the inversion layer electrons is shown in Fig. 3.1. In order to analyze the electron distribution in the inversion layer analytically, it is convenient to approximate this well with a triangular well. The slope of the triangular well's wall is the effective electric field F_{EFF}[1], which by definition is the average electric field 'felt' by the inversion layer electrons.

$$F_{EFF} = \frac{\int_{-\infty}^{+\infty} F(x)\rho(x)dx}{\int_{-\infty}^{+\infty} \rho(x)dx} \tag{3.2}$$

where $\rho(x)$ is the electron density. We shall use Poisson equation $\rho(x)=\varepsilon F(x)'$ and note that the integral limits can be made to coincide with the top and bottom of the inversion layer to obtain

$$F_{EFF} = \frac{\int F(x)F'(x)dx}{\int F'(x)dx} = \frac{(F_t^2 - F_b^2)/2}{F_t - F_b} = \frac{(F_t + F_b)}{2}. \tag{3.3}$$

Here F_t and F_b represent the magnitude of the electric field at the top and the bottom of the inversion layer. Thus, F_{EFF} turns out to simply be the arithmetic average of the top and bottom electric fields. F_{EFF} can also be expressed in terms of the inversion and depletion charges and the dielectric permittivity of silicon [3.7]:

$$F_{EFF} = \frac{F_t + F_b}{2} = \frac{(Q_{DEP} + Q_{INV}) + Q_{DEP}}{2\varepsilon_{Si}} = \frac{2Q_{DEP} + Q_{INV}}{2\varepsilon_{Si}} \tag{3.4}$$

---

[1] Letter F is used in this chapter to denote the electric field to distinguish it from the energy levels. In the next chapter, we shall revert to a more traditional notation: E.

In the above derivation, we assumed that charge density of the inversion electrons is much greater then that of the depletion charge, and therefore the derivation is only valid in the strong inversion regime.

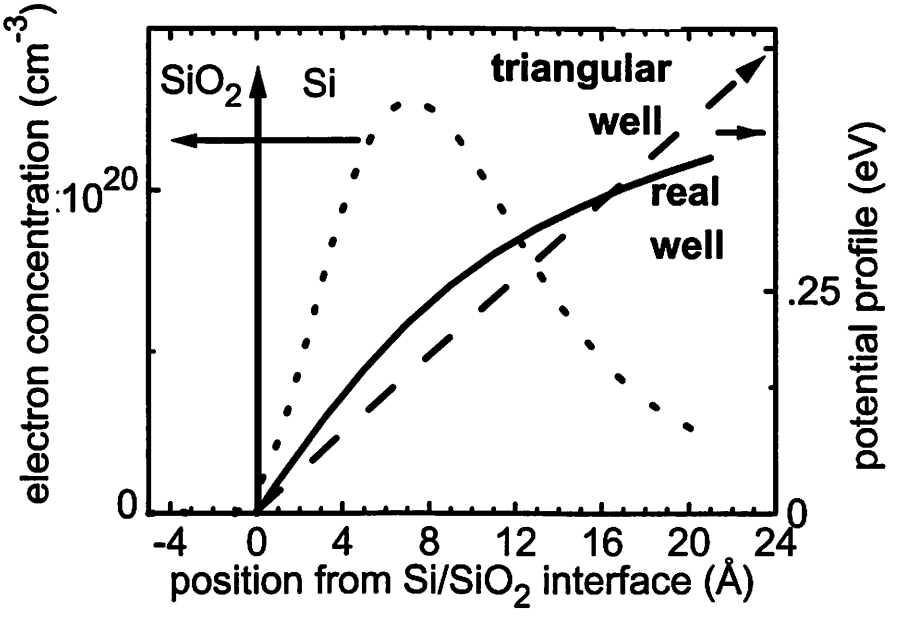


Fig. 3.1. Comparison between the real potential well confining the inversion electrons and the triangular model well.

3.3.2 Electron energy, charge centroid, and impingement frequency

Electron wave function in the triangular potential well is described by Schrödinger equation:

$$\frac{\hbar^2}{2}\frac{d}{dx}\left(\frac{1}{m}\frac{d\Psi}{dx}\right)+(E+Fx)\Psi=0. \qquad (3.5)$$

If x is replaced with a dimensionless variable [3.8]

$$\xi = \left(x + \frac{E}{qF} \right) \left(\frac{2mqF}{\hbar^2} \right)^{1/3}, \tag{3.6}$$

Schrödinger equation takes on a very simple form:

$$\Psi'' + \xi \Psi = 0. \tag{3.7}$$

The solution to this equation is the Airy function of a single dimensionless variable ξ, and therefore the scaling of all the parameters related to the electron distribution with changing electric field F is given by equation (3.6). So that the lateral spread of the wavefunction and charge centroid are proportional to

$$x_C \propto \left(\frac{2mqF}{\hbar^2} \right)^{-1/3}. \tag{3.8}$$

The inverse of x_C, the characteristic k-vector of the wavefunction is

$$k \propto \left(\frac{2mqF}{\hbar^2} \right)^{1/3}. \tag{3.9}$$

Energy scales according to

$$E \propto qFx_C \propto qF \left(\frac{2mqF}{\hbar^2} \right)^{-1/3}, \tag{3.10}$$

and the impingement frequency according to

$$f_{IMP} \propto \frac{E}{\hbar} \propto (qF)^{2/3} (2m\hbar)^{-1/3}. \tag{3.11}$$

3.3.3 Impingement frequency – numerical analysis

The theory presented in the previous section demonstrates that the impingement frequency is a function of the effective electric field only. In this section, we shall use the results of numerical simulations to confirm that conclusion. In addition, the numerical

simulations take into account the presence of both heavy and light electrons and can therefore serve as the basis for a more precise model than the one given by (3.11).

By numerically solving Schrödinger and Poisson equations self-consistently [3.9] we can determine the electron density Ψ as a function of x at any gate bias. The electron density near the dielectric interface (x=0) can be approximated by

$$A\sin(kx) = \frac{A\exp(ikx) - A\exp(-ikx)}{2} \tag{3.12}$$

The electron flux impinging on the dielectric interface is then given by

$$j = \frac{i\hbar}{8m_1} A\exp(-ikx)\frac{d}{dx}(A\exp(ikx)) = \frac{i\hbar A^2}{8m_1}k \tag{3.13}$$

Finally, f_{IMP} is given by the ratio of the flux j and Q_{INV}. The results of this numerical solutions for MOS transistors with different oxide thicknesses and substrate doping concentrations are presented in Figure 3.2. As expected, the impingement frequency depends only on the effective electric field. Furthermore, this dependence closely follows the $F^{3/2}$ trend predicted by equation (3.11). An even better empirical fit can be achieved through the polynomial function shown in Figure 3.2.

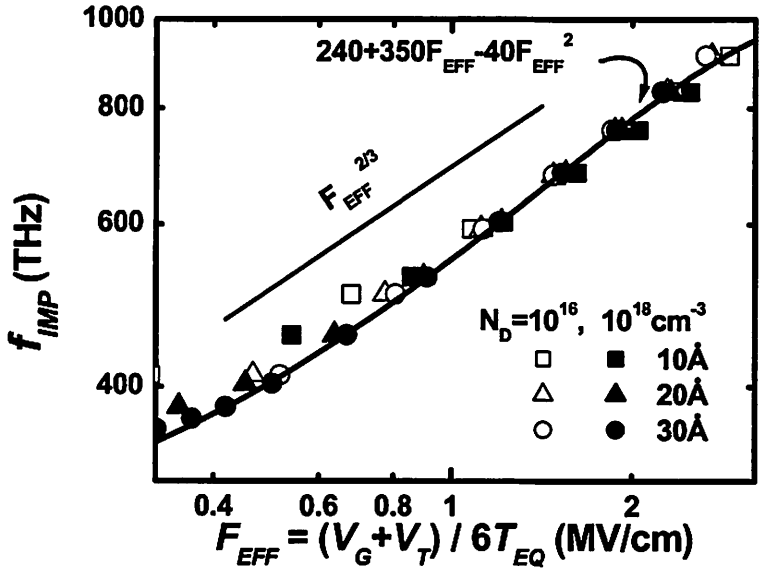


Fig. 3.2 Electron impingement frequency f_{IMP} is a function of F_{EFF} only. As described in the text, f_{IMP} is determined from the inversion layer electron distribution obtained from QM simulation [3.9].

3.4 Tunneling through multiple layer dielectric layers

In this section, we shall describe how the electron transmission probability (T) through a multi-layer potential barrier (Fig 3.3) is calculated. We shall consider a plane wave e^{-ik_1x} impinging on the barrier, reflected wave re^{ik_1x}, and the transmitted wave te^{-ik_2x}. (Please note that in general the wave numbers k_1 and k_2 as well as the effective electron masses m_1 and m_2 are different in the cathode and anode.) Transmission probability T by definition is equal to $|t|^2$. The magnitude of the transmitted wave t is equal to the product of transition coefficients through all three interfaces shown in Figure 3.3 and the attenuation coefficients in the two classically forbidden regions $\mathcal{A}$ and $\mathcal{B}$.

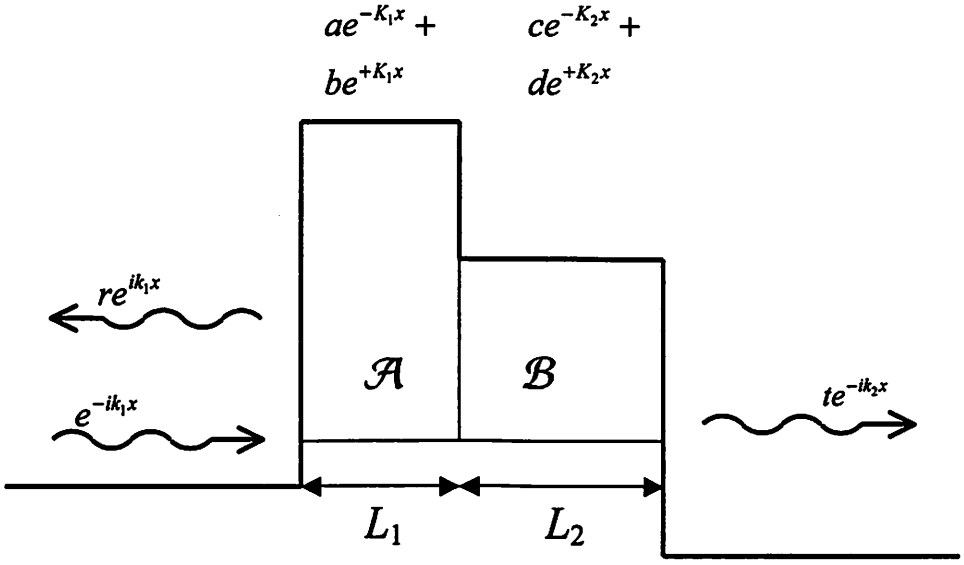


Fig. 3.3. Diagram of a double-layer potential barrier.

3.4.1 Electron transmission through semiconductor/dielectric and dielectric/dielectric interfaces

Here we shall calculate the wave transmission coefficient through the simplest kind of interface when the electron energy is high enough so that the regions on both sides of the interface are classically allowed (Fig. 3.4).

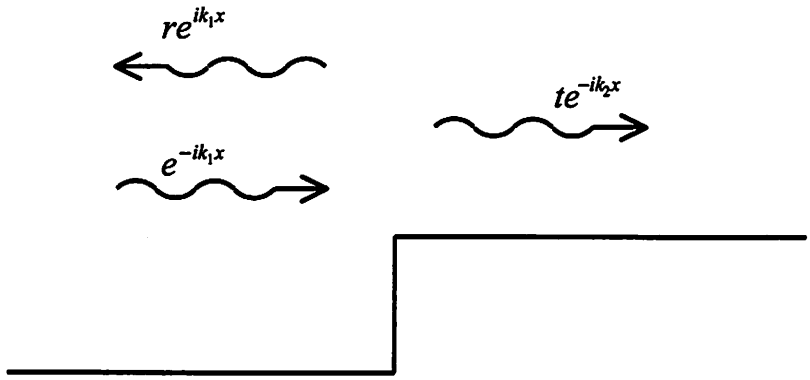


Fig. 3.4. Diagram of electron reflection and transmission at an interface between two semiconductor regions.

The wave function on the left side of the interface, once again, is given by $e^{-ik_1x} + re^{ik_1x}$, and on the right side by te^{-ik_2x}. According to the Schrödinger equation (3.5) the wave function as well as its derivative divided by the effective electron mass should be continuous across the interface yielding:

$$\begin{cases} 1+r=t \\ (-1+r)\dfrac{k_1}{m_1} = -t\dfrac{k_2}{m_2} \end{cases} \tag{3.14}$$

Solving these equations we obtain $t = \dfrac{2}{1+\dfrac{k_2}{k_1}\dfrac{m_1}{m_2}}$ or

$$T = \frac{4}{\left(1+\dfrac{k_2}{k_1}\dfrac{m_1}{m_2}\right)^2}. \tag{3.15}$$

The calculation of the transmission coefficient for the other cases is more cumbersome; details are provided in Appendix 3.A. However, the result (3.15) also holds when the regions on both sides of the interface are classically forbidden. When the region on one of

the sides of the interface is allowed and the region on the other side is forbidden, the expression is slightly modified to become

$$T = \frac{4}{1 + \left(\frac{k_2}{k_1}\frac{m_1}{m_2}\right)^2}. \tag{3.16}$$

3.4.2 WKB approximation

While the transmission coefficients through various interfaces calculated in 3.4.1 are important in determining the precise value of tunneling probabilities, they are essentially constant and independent of either gate bias or dielectric layers' thickness, and therefore have a very limited impact on the study of the gate dielectric scaling. A much more profound effect can be attributed to the amount of wave function attenuation, A, within the classically forbidden dielectric regions. There are several possible ways to calculate A. First, Schrödinger equation can be solved numerically within each dielectric region. Second, as long as the dielectric barriers have trapezoidal shape, the result can be expressed in terms of the Airy functions. Third, an approximate analytical solution to Schrödinger equation (3.5) can be used. Only the latter approach provides a simple analytical expression for the amount of wave function attenuation. The Wentzel-Kramers-Brillouin (WKB) approximation (also known as quasi-classical approximation) has been the method of choice for calculation tunneling probabilities. According to the approximation [3.10]

$$A = \exp\left(-2\int_0^L \sqrt{\frac{2m_d}{\hbar^2}(V(x)-E)}dx\right) \Big/ \sqrt{\frac{V(L)-E}{V(0)-E}} \tag{3.17}$$

Here L is the thickness of the dielectric, m_d is electron effective mass, $V(x)$ is the potential barrier height as a function of the position, and E the electron energy. When the integral is evaluated for a dielectric with barrier height ϕ_B, and voltage V_{OX} applied across it the expression becomes.

$$\exp\left[-4L/3\hbar V_{OX}\sqrt{2mq}\cdot\phi_B^{3/2}\left(1-\left(1-V_{ox}/\phi_B\right)^{3/2}\right)\right]\Big/\sqrt{1-V_{ox}/\phi_B} \qquad (3.18)$$

(E is assumed negligible compared to ϕ_B.) The first exponential term has been traditionally used to model direct tunneling. However, such a simple model becomes rather inaccurate as V_{OX} approaches ϕ_B (Fig. 3.5). The second order square-root term adds little complexity, while greatly enhancing the model's precision (Fig. 3.5). This new second-order WKB approximation for the tunneling probability provides a very good match to the precise numerical solution for a variety of gate dielectrics and the entire range of the gate biases corresponding to the direct tunneling regime.

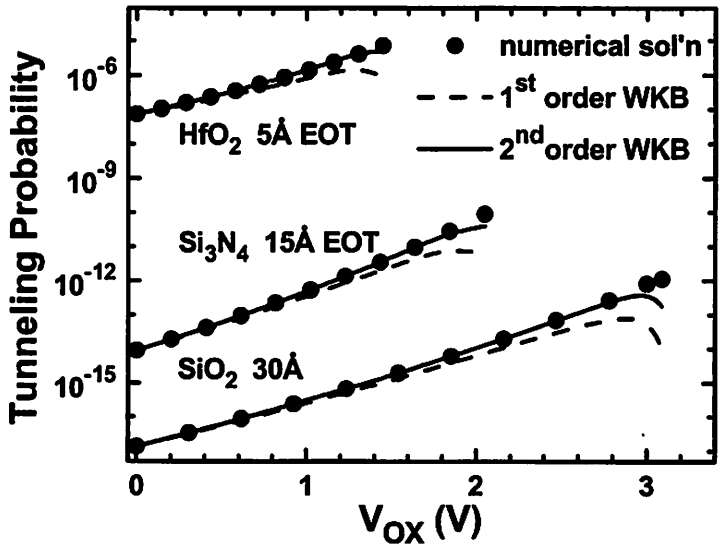


Fig. 3.5. 2nd order WKB model gives an excellent approximation to the direct tunneling probability.

3.5 Scaling of single layer gate dielectrics

The general approach to the electron tunneling described by (3.1) should be equally valid for both single- and multi-layer gate dielectrics. Therefore, we shall start with a simpler single-layer case, and then generalize our findings to double-layer gate dielectrics.

3.5.1 Analytical model validation for single layer dielectrics

The validity of the simple analytical model (3.1) can be verified by comparing its results to the experimentally measured tunneling current through various single-layer dielectrics. The two most commonly used dielectrics to date are silicon oxide and silicon nitride. Other (high-κ) gate dielectrics have been fabricated recently. However, the values of the barrier height and the effective electron mass for these materials have not yet been clearly established. In addition during the fabrication of MOS structures with high-κ dielectrics interfacial layers of uncertain composition are usually formed. Therefore, we only use the experimental results for single-layer SiO_2 and Si_3N_4 to validate the analytical model. According to Figure 3.6, our analytical model correctly predicts the magnitude and the general shape of the leakage current curve.

3.5.2 Tunneling attenuation coefficient

An understanding of the gate leakage current dependence on the equivalent gate oxide thickness (t_{eq}) a key issue in gate dielectric scaling. Fortunately, this dependence comes primarily through the exponential function described in equation (3.18), and

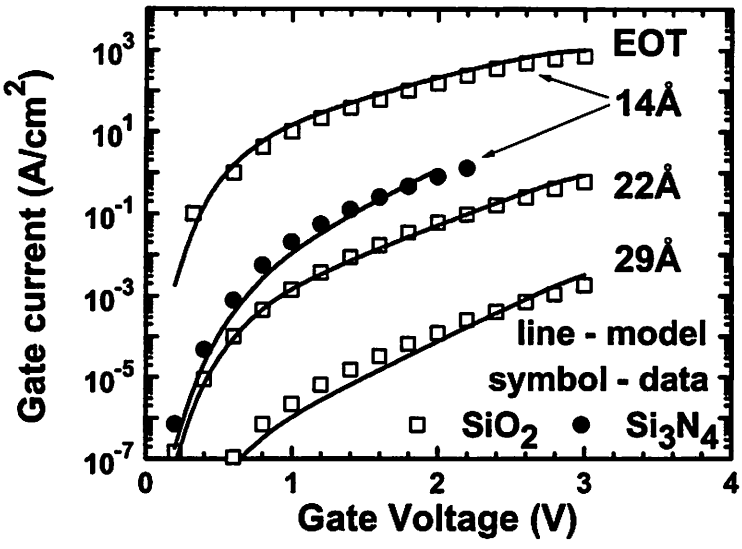


Fig. 3.6. The simple analytical model for direct tunneling provides a close match to the experimental data from [3.11-3.13].

therefore, gate leakage current can be approximately expressed as

$$J_G \sim exp(-\alpha t_{eq}). \tag{3.19}$$

Coefficient α is given by

$$\alpha = \left(4\kappa_D / 3\hbar V_{OX} \kappa_{OX}\right) \sqrt{2mq}\,\phi_B^{3/2} \left(1 - \left(1 - V_{OX}/\phi_B\right)^{3/2}\right), \tag{3.20}$$

and has the dimensions of inverse length. Thus we shall refer to α as "tunneling attenuation coefficient." According to equation (3.19), the leakage current values for various dielectrics are confined to a series of universal straight lines as shown in Figure 3.7. The slope of these lines α is a function of the dielectric parameters and V_{OX}. High-κ dielectrics have larger values of α and therefore smaller leakage current.

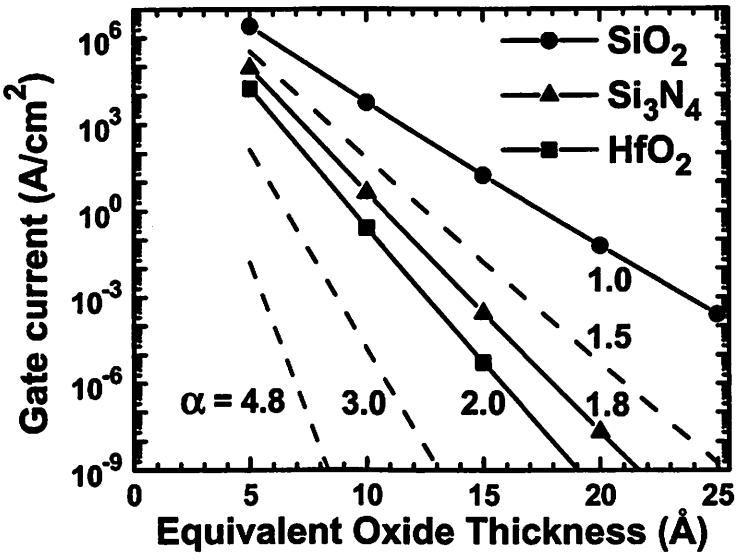


Fig. 3.7. Leakage currents for various single-layer dielectrics falls on a universal family of lines. Each line is essentially determined by a single number – tunneling attenuation coefficient α.

3.6 Scaling of double-layer gate dielectrics

In the previous section, it has been shown that as far as the scaling of single-layer dielectrics is concerned the leakage current can be expressed by a very simple expression (3.19). Here we intend to show that an equation of the same form can describe the scaling of double-layer dielectrics as well. Figure 3.8 shows the calculated transmission probability for a stack comprised of two dielectric layers (SiO_2 and HfO_2). The five lines represent the five cases in which HfO_2 contributes 0%, 25%, 50%, 75% and 100% of the equivalent oxide thickness of the entire stack. Since a straight line corresponds to each of these cases, we realize that the tunneling probability through a double-layer stack can also be effectively described by the tunneling attenuation coefficient α.

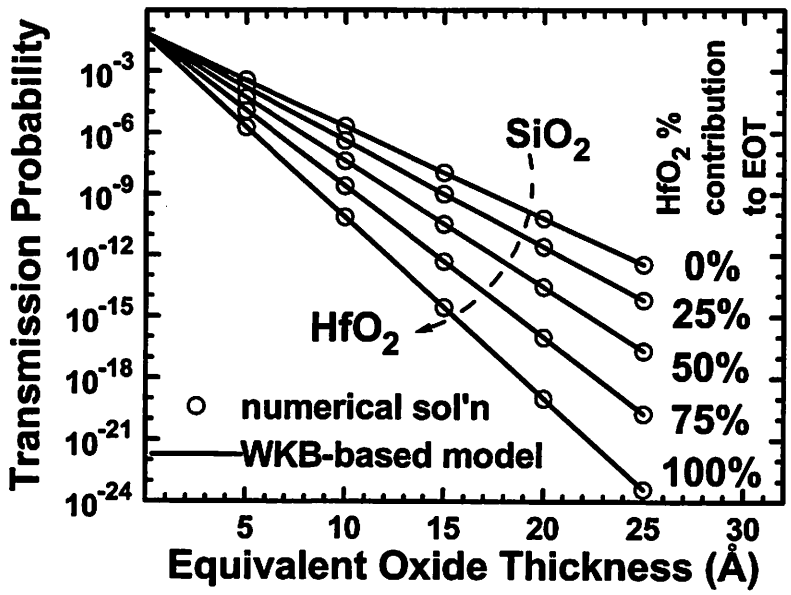


Fig. 3.8. Leakage currents for various dielectric stacks fall on the same family of lines as for single layer dielectrics.

3.6.1 Tunneling attenuation coefficient for double layer stacks

We shall now determine how the tunneling attenuation coefficient α for a double-layer dielectric depends on the composition of the dielectric stack. The simplest way to obtain α would be to assume that it is determined by a weighted average of the tunneling attenuation coefficients of the two dielectrics (α_1 and α_2): $\alpha \approx \alpha_1 f_1 + \alpha_2 f_2$. ($f_1$ and f_2 are the fractional contribution to the total EOT from the two layers, $f_1+f_2=1$) In reality however, α can be better approximated by a quadratic relation (Fig 3.7):

$$\alpha \approx \alpha_1 f_1 + \alpha_2 f_2 + c f_1 f_2 . \tag{3.21}$$

The curvature of the lines in Figure 3.7, c is also a function of the dielectric parameters and V_{OX}, and is approximately given by

$$c \approx {V_{ox}}/{3\hbar}\left(\kappa_1\sqrt{{qm_1}/{2\phi_1}} - \kappa_2\sqrt{{qm_2}/{2\phi_2}}\right) \qquad (3.22)$$

The details of the derivation of (3.22) are given in Appendix 3.B. As shown in Figure 3.7, the results of the equations (3.21) and (3.22), shown by solid and dashed lines, closely match the results of numerical solution (shown by symbols). It is important to note that the sign of c depends on the order of the dielectrics in the stack while its magnitude remains the same, as predicted by (3.22). Generally, c is positive, and the tunneling current is lower when the high-κ layer is on the cathode (emitting) side. This conclusion can also be draw intuitively if we examine the dielectric band diagrams shown in Figure 3.7 (the same gate bias is assumed in both cases). According to the WKB approximation (3.17) the gate leakage current, roughly speaking, is determined by the area under the dielectric barrier. The area under the dielectric band diagram shown with dashed line (tunneling from the high-κ side) is clearly larger, and therefore lower leakage current is expected.

3.7 Conclusion

A simple analytical model for direct tunneling through gate dielectric stacks is introduced. The model is based on a refinement of the WKB approximation, and corresponds closely to the experimental and simulation results.

According to this model, in a dual-layer dielectric, the tunneling current is lower when the high-κ layer is on the emitting side (the cathode for electron tunneling), compared to the case when SiO_2 layer is on the emitting side.

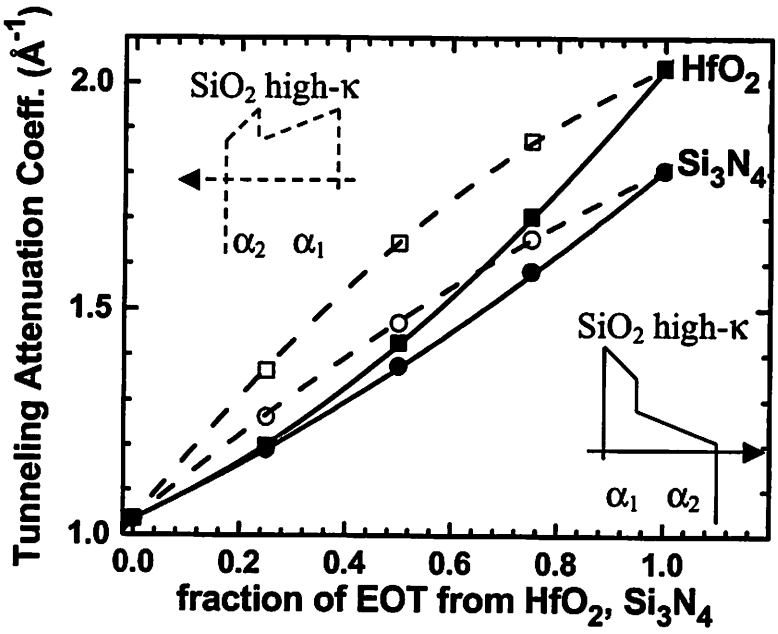


Fig. 3.9. A stack has intermediate values of tunneling attenuation coefficients that are determined by the composition of the stack and the polarity of tunneling current.

In order to simplify the prediction of the gate leakage current, we have introduced the notion of the tunneling attenuation coefficient, which is a simple function of the dielectric parameters and gate voltage for both single- and multi-layer stacks. This coefficient alone allows for accurate prediction of the gate leakage current as a function of the equivalent oxide thickness. Thus the tunneling attenuation coefficient is an important parameter governing the scalability of a dielectric stack; and one single family of universal lines define the realm of leakage current through single- and multi-layer high-κ gate dielectric stacks.

3.8 References

3.1 W.-C. Lee, 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, No. 7, pp. 1366-1373, July 2001.

3.2 Yee Chia Yeo, Qiang Lu, Wen Chin Lee, Tsu-Jae King, Chenming Hu, Xiewen Wang, Xin Guo, T.P. Ma, "Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric," *IEEE Electron Device Letters*, Vol. 21, No. 11, pp. 540-542, Nov. 2000.

3.3 Ying Shi, Xiewen Wang, T. P. Ma, "Tunneling leakage current in ultrathin (<4 nm) nitride/oxide stack dielectric," *IEEE Electron Device Letters*, Vol. 19, No. 10, pp. 388-390, Oct. 1998.

3.4 Nian Yang, W. Kirklen Henson, John R. Hauser, Jimmie J. Wortman, "Modeling study of ultrathin gate oxides using direct tunneling current and capacitance-voltage measurements in MOS devices," *IEEE Transactions on Electron Devices*, Vol. 46, No. 7, pp. 1464-1471, July 1999.

3.5 J. Zhang, J. S. Yuan, Y. Ma, A. S. Oates, "Design optimization of stacked layer dielectrics for minimum gate leakage currents," *Solid State Electronics*, Vol. 44, pp. 2165-2170, Dec. 2000.

3.6 S. Mudanai, Yang-Yu Fan, Qiqing Ouyang, A. F. Tasch, S. K. Banerjee, "Modeling of direct tunneling current through gate dielectric stacks," *IEEE Transactions on Electron Devices*, Vol. 47, No. 10, pp. 1851-1857, Oct. 2000.

3.7 Dragica Vasileska, David K. Ferry, "Scaled silicon MOSFET's: universal mobility behavior," *IEEE Transactions on Electron Devices*, Vol. 44, No. 4, pp. 577-583, Apr. 1997.

3.8 L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, 3rd ed., Pergamon Press, Oxford, New York, 1991.

3.9 available www-device.eecs.berkeley.edu/qmcv.html

3.10 Jon Mathews, R. L. Walker, Mathematical methods of physics, 2nd addition, Addison–Wesley Publishing Company, Inc., New York, p. 27, 1970.

3.11 Yee Chia Yeo, Qiang Lu, Wen Chin Lee, Tsu-Jae King, Chenming Hu, Xiewen Wang, Xin Guo, T.P. Ma, "Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric," *IEEE Electron Device Letters*, Vol. 21, No.11, pp. 540-542, Nov. 2000.

3.12 S.-H. Lo, D. A. Buchanan, Y. Taur, W. Wang, "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET's," *IEEE Electron Device Letters*, Vol. 18, No. 5, pp .209-211, May 1997.

3.13 S. Song, J. H. Yi, W. S. Kim, J. S. Lee, K. Fujihara, H. K. Kang, J. T. Moon, M. Y. Lee, "CMOS device scaling beyond 100 nm," *Tech. Dig. International Electron Devices Meeting*, pp. 235-238, 2000

Appendix 3.A

Here we are going to analytically calculate the tunneling probability through the double rectangular barrier shown in Figure 3.3. As discussed in section 3.4.1 the electron wave function in the cathode and anode are given by $e^{-ik_1x} + re^{ik_1x}$ and te^{-ik_1x} respectively. The decaying wave function in the dielectric region $\mathcal{A}$ is given by a superposition of two linearly independent solutions to Schrödinger equation $ae^{-K_1x} + be^{+K_1x}$, where K_1 is the "imaginary wave vector" within the dielectric layer $\mathcal{A}$. The value of K_1 is determined by dielectric barrier height ϕ_{B1}, effective electron mass m_d and electron energy E: $K_1 = \sqrt{\frac{2m_{d1}}{\hbar^2}(\phi_{B1} - E)}$. Similarly, the wave function in the dielectric region $\mathcal{B}$ is given by $ce^{-K_2x} + de^{+K_2x}$.

As discussed in section 3.4.1, three sets of boundary conditions have to be satisfied (at x=0, x=L_1, and x=L_1+L_2). This leads to the following set of 6 linear equations with 6 variables a, b, c, d, r, and t:

$$\begin{cases} 1+r=a+b & \text{(A.1)} \\ i(-1+r)\dfrac{k_1}{m_1}=(-a+b)\dfrac{K_1}{m_{d1}} & \text{(A.2)} \\ ae^{-K_1L_1}+be^{K_1L_1}=ce^{-K_2L_1}+de^{K_2L_1} & \text{(A.3)} \\ \left(-ae^{-K_1L_1}+be^{K_1L_1}\right)\dfrac{K_1}{m_{d1}}=\left(-ce^{-K_2L_1}+de^{K_2L_1}\right)\dfrac{K_2}{m_{d2}} & \text{(A.4)} \\ ce^{-K_2(L_1+L_2)}+de^{K_2(L_1+L_2)}=te^{-ik_2(L_1+L_2)} & \text{(A.5)} \\ \left(-ce^{-K_2(L_1+L_2)}+de^{K_2(L_1+L_2)}\right)\dfrac{K_2}{m_{d2}}=-ite^{-ik_2(L_1+L_2)}\dfrac{k_2}{m_2} & \text{(A.6)} \end{cases}$$

To simplify future expressions we introduce the following parameters:

$$A=e^{K_1L_1},\quad B=e^{K_2L_2},\quad C=e^{K_2(L_1+L_2)},\quad D=e^{K_2L_1},\quad \gamma=\left(\frac{K_2}{m_{d2}}+i\frac{k_2}{m_2}\right)\Big/\left(\frac{K_2}{m_{d2}}-i\frac{k_2}{m_2}\right)$$

A convenient way to solve these equations is to combine the first and second equations to eliminate the variable r, and combine the fifth and sixth equations to eliminate variable t, resulting in:

$$\begin{cases} 2=a\left(1+\dfrac{K_1m_1}{im_{d1}k_1}\right)+b\left(1-\dfrac{K_1m_1}{im_{d1}k_1}\right) & \text{(A.7)} \\ c=dC^2\gamma & \text{(A.8)} \end{cases}$$

Next, we substitute (A.8) into (A.3) and (A.4), use the resulting two equations to eliminate d, and then use that result in conjunction with (A.7) to eliminate a. Thus we arrive at

$$b=2\cdot\left[1-\frac{K_1m_1}{im_{d1}k_1}+\left(1+\frac{K_1m_1}{im_{d1}k_1}\right)A^2\frac{D^2+\gamma C^2-\dfrac{K_1m_{d2}}{K_2m_{d1}}\left(D^2-\gamma C^2\right)}{D^2+\gamma C^2+\dfrac{K_1m_{d2}}{K_2m_{d1}}\left(D^2-\gamma C^2\right)}\right]^{-1}$$

Now we can find all the previously eliminated variables including t, and by finding the square of the absolute value of t, we find the transition probability T. The full expression

for T is very long, so that in order to keep it as compact as possible we have removed all the effective masses which always accompany the corresponding k-vectors. (For example, one should read k_1/m_1 instead of k_1, or K_1/m_{d1} instead of K_1.)

$$
\begin{aligned}
T = {} & (64\,A^2 D^2 C^2 k_1^2 K_1^2 K_2^2)/ \\
& (D^4 (k_2^2 + K_2^2)(k_1^2 ((1 + A^2) K_1 + (1 - A^2) K_2)^2 + K_1^2 ((1 - A^2) K_1 + (1 + A^2) K_2)^2) + \\
& C^4 (k_2^2 + K_2^2)(k_1^2 ((1 + A^2) K_1 + (-1 + A^2) K_2)^2 + K_1^2 ((-1 + A^2) K_1 + (1 + A^2) K_2)^2) - \\
& 2 D^2 C^2 (-16 A^2 k_1 k_2 K_1^2 K_2^2 + k_1^2 (k_2^2 - K_2^2)((1 + A^2)^2 K_1^2 - (-1 + A^2)^2 K_2^2) + \\
& K_1^2 (k_2^2 - K_2^2)((-1 + A^2)^2 K_1^2 - (1 + A^2)^2 K_2^2)))
\end{aligned}
$$

This is a precise albeit very complicated expression. However, if we realize that as long as the dielectric layers are at least a couple of angstroms thick the parameters A, B, C, and D are all much greater than 1. The expression can then be drastically simplified:

$$
T = \frac{64A^2D^2C^2\left(\frac{k_1}{m_1}\right)^2\left(\frac{K_1}{m_{d1}}\right)^2\left(\frac{K_2}{m_{d2}}\right)^2}{A^4C^4\left(\left(\frac{k_1}{m_1}\right)^2+\left(\frac{K_1}{m_{d1}}\right)^2\right)\left(\left(\frac{k_2}{m_2}\right)^2+\left(\frac{K_2}{m_{d2}}\right)^2\right)\left(\frac{K_1}{m_{d1}}+\frac{K_2}{m_{d2}}\right)^2}
$$

The expression can be further simplified since D=BC. We shall also break down the expression into a product of several terms so that the significance of these terms can be easily understood.

$$
T = \frac{4}{1+\left(\frac{K_1 m_1}{k_1 m_{d1}}\right)^2}\cdot\frac{1}{A^2}\cdot\frac{4}{\left(1+\frac{K_2 m_{d1}}{K_1 m_{d2}}\right)^2}\cdot\frac{1}{B^2}\cdot\frac{4}{1+\left(\frac{k_2 m_{d2}}{K_2 m_2}\right)^2}
$$

The first term represents the transmission probability at the interface between the cathode and the first dielectric layer (Fig 3.3). The second term represents the attenuation in the first dielectric layer. The third term represents the transmission probability at the interface between the two dielectrics (compare with expression (3.16)). The fourth term represents the attenuation in the second dielectric layer. Finally, the last term represents

the transmission probability at the interface between the second dielectric layer and the cathode.

Appendix 3.B

In order to derive the quadratic approximation (3.22) we shall consider a gate stack comprised of two dielectric layers with the fractional contributions to the total equivalent oxide thicknesses of f and 1-f respectively. If the voltage applied across the entire stack is V_{OX} then the voltage drop across the first layer is fV_{OX} and the WKB-based tunneling probability through the first layer is

$$\exp\left[-4t_{OX}/3\hbar V_{OX}\,\frac{\kappa_1}{\kappa_{OX}}\sqrt{2m_1 q}\cdot\left(\phi_{B1}^{3/2}-(\phi_{B1}-fV_{OX})^{3/2}\right)\right].$$

Similarly, the tunneling probability through the second layer is

$$\exp\left[-4t_{OX}/3\hbar V_{OX}\,\frac{\kappa_2}{\kappa_{OX}}\sqrt{2m_2 q}\cdot\left((\phi_{B2}-fV_{OX})^{3/2}-(\phi_{B2}-V_{OX})^{3/2}\right)\right]$$

The total tunneling probability expressed as the product of the previous two expressions can be written as $\exp[-\alpha t_{OX}]$, where

$$\alpha=-\frac{4\sqrt{2q}}{3\hbar V_{OX}\kappa_{OX}}\left[\kappa_1\sqrt{m_1}\left(\phi_{B1}^{3/2}-(\phi_{B1}-fV_{OX})^{3/2}\right)+\kappa_2\sqrt{m_2}\left((\phi_{B2}-fV_{OX})^{3/2}-(\phi_{B2}-V_{OX})^{3/2}\right)\right].$$

When we perform the series expansion in powers of f the second order term in the expansion is going to give us the quadratic coefficient c in equation (3.21).

$$c=V_{ox}/3\hbar\left(\kappa_1\sqrt{qm_1/2\phi_1}-\kappa_2\sqrt{qm_2/2\phi_2}\right)$$

Chapter 4 Channel Mobility in MOSFETs with Alternative Gate Dielectrics

4.1 Introduction - Universal mobility model for SiO_2

Carrier channel mobility is one of the most important physical parameters that determine MOSFET performance. Carrier mobility at Si-SiO_2 interface depends on several factors: the crystalline orientation of Si surface [4.1], channel doping concentration N_D, gate bias V_G, and gate oxide thickness T_{OX}. It is interesting to note that mobility does not usually depend on the way SiO_2 is grown (provided reasonable care is taken to insure a clean interface). Mobility dependence on the crystalline orientation is also not an issue for the present-date CMOS technology, since only (100) Si wafers are now used for integrated circuit fabrication. Furthermore, channel carrier mobility does not depend on each of the remaining 3 parameters (N_D, V_G, T_{OX}) independently, but instead depends on the combination of these three parameters. This combination of the three parameters is the effective electric field E_{EFF} that has been introduced in the previous chapter (equation (3.3)).

The fact that mobility depends only on E_{EFF} was first reported by Sabnis and Clemens [4.2]. The functional dependence of mobility on E_{EFF} is known as the universal mobility model. The physical foundation of this model will be discussed in the subsequent sections. Here we shall only seek to emphasize that this model has been extremely successful, because of its applicability to all Si MOS transistors developed over the past two decades.

As we have already discussed in the previous chapters, it is likely that over the next decade, SiO_2 will be replaced as the gate dielectric by a high-κ material. Consequently the universal mobility model, in its present form, is not going to be applicable in most cases.

We shall start this chapter by reexamining the physics behind the universal mobility model for (100) Si-SiO_2 interface. We will then propose an alternative model to account for the degradation of mobility at high E_{EFF}. This model is in a better agreement with experimental results then the commonly used surface roughness model. The new model also explains why a lower mobility is observed for high-κ dielectrics.

4.2 Electron scattering mechanisms

There are three important mechanisms that determine electron channel mobility (Fig. 4.1). First is Coulombic scattering by the ionized dopants in the channel region, interface states, and the bulk charges in the gate dielectric. This mechanism is dominant at low gate biases. (At higher gate biases, as the magnitude of the mobile inversion charge is increased, the Coulombic centers are screened and the amount of Coulombic scattering is dramatically reduced.) The second mechanism is phonon scattering, which

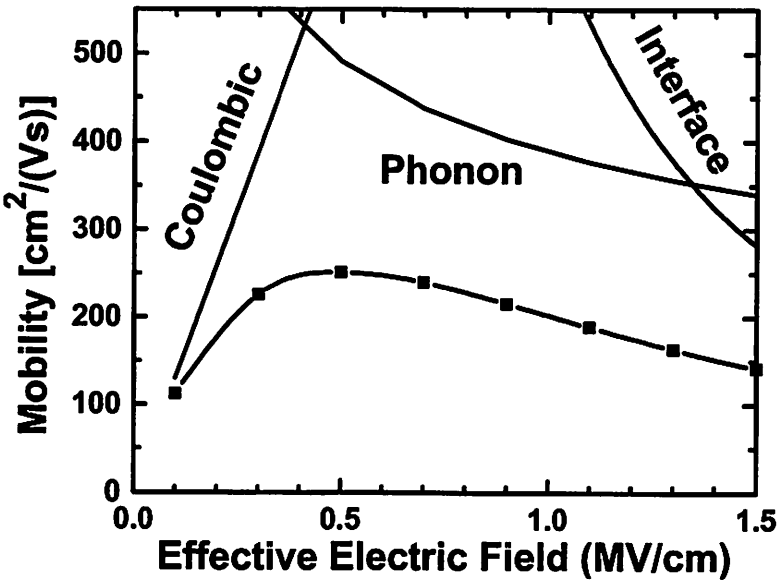


Fig. 4.1. Combination of three scattering mechanisms determines the surface channel mobility.

dominates at the intermediate gate biases. The third mechanism is related to the scattering at the dielectric semiconductor interface and is dominant at high gate biases. The effects of all three mechanisms can be combined using Mathiessen's rule:

$$\frac{1}{\mu} = \frac{1}{\mu_{COUL}} + \frac{1}{\mu_{PH}} + \frac{1}{\mu_{INT}} \tag{4.1}$$

The traditional universal mobility model that we introduced in the previous chapter represents the combination of the last two terms of (4.1) and is applicable in the intermediate to high range of E_{EFF}.

$$\frac{1}{\mu_{UNIV}} = \frac{1}{\mu_{PH}} + \frac{1}{\mu_{INT}}. \tag{4.2}$$

The performance of digital CMOS circuits depends primarily on the amount of current that a transistor can supply under high gate bias. Thus it is especially important to

understand and model the carrier mobility at high E_{EFF}. In the following sections, we shall only briefly discuss the Columbic and phonon scattering, and instead focus our attention on the high E_{EFF} region.

4.2.1 Coulombic scattering

The task of calculating the rate of carrier scattering by Coulombic charges in the vicinity Si-SiO_2 interface has been addressed by several researchers [4.3-4.7]. The calculations of scattering rate in the electric quantum limit are rather complex; therefore we shall consider only the phenomenological results of those calculations. The scattering rate is clearly dependent of the density of scattering centers and their location in the gate dielectric. The closer the traps are to the channel, the higher the scattering rate (Fig. 4.2). The scattering rate also depends on the amount of the inversion charge. A larger inversion charge density provides more screening of the charge centers and reduces the scattering rate. For charge centers located far from the inversion layer the amount of screening depends linearly on the inversion charge density N_{INV} (Fig. 4.2). For the scattering centers located closer to the inversion layer the screening dependence is sub-linear. In general, Coulombic mobility as a function of the inversion charge density can be successfully modeled by a combination of a linear and a sub-linear terms.

$$\frac{1}{\mu_{COUL}} = \frac{1}{AN_{INV}} + \frac{1}{BN_{INV}^{0.6}} \tag{4.3}$$

The first term corresponds to the scattering by the bulk charges in the dielectric; the second term corresponds to the scattering by the interface charges. Coefficients A and B depend on the amount of trapped charge.

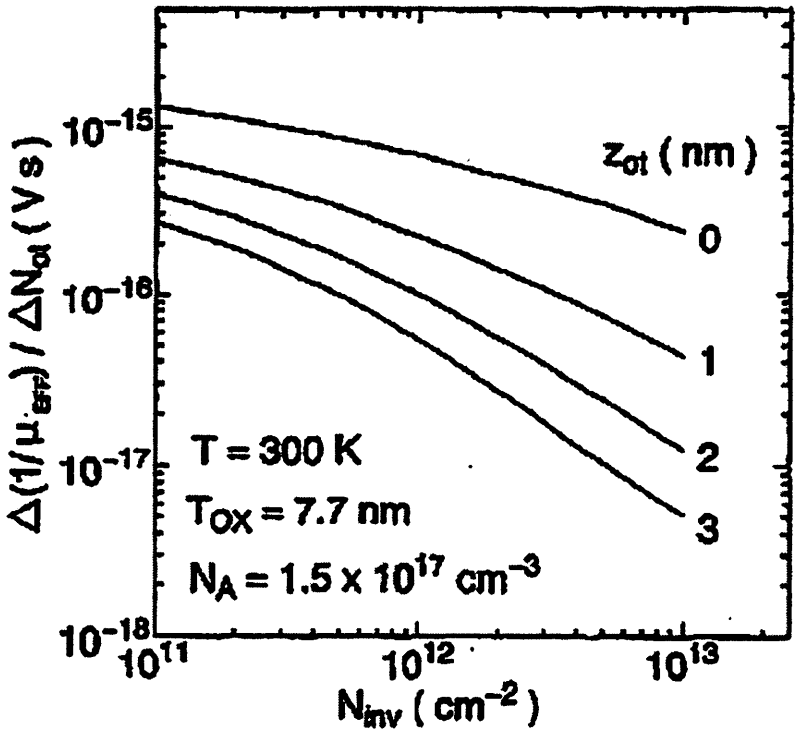


Fig. 4.2. This plot from reference [4.6] shows the calculated scattering rate per unit Coulombic charge as a function of the inversion charge density for various positions of the scattering charge within the gate dielectric.

4.2.2 Phonon scattering

Phonon scattering is the first of the two components of the universal mobility model (4.2). There are several types of lattice vibrations that are responsible for inversion layer electron scattering: acoustic phonons, optical phonons and the interface optical modes induced by the proximity of the polar SiO_2 layer [4.8]. In addition, one has to take into account the fact that the Si conduction band in the inversion layer is split due to quantum confinement into a number of sub-bands, so that multiple intra-sub-band and inter-sub-band scattering events must be taken into account. Therefore, the only way to

precisely determine the phonon contribution to the universal mobility model is through a computer simulation [4.8]. A result of such Monte Carlo simulations is shown in Figure 4.3. This result predicts that μ_{PH} is proportional to E_{EFF} to the power -0.3. Once again, we are not going to describe the detail of phonon scattering, restricting our discussion to a simple argument that correctly predicts the dependence of phonon scattering on E_{EFF}.

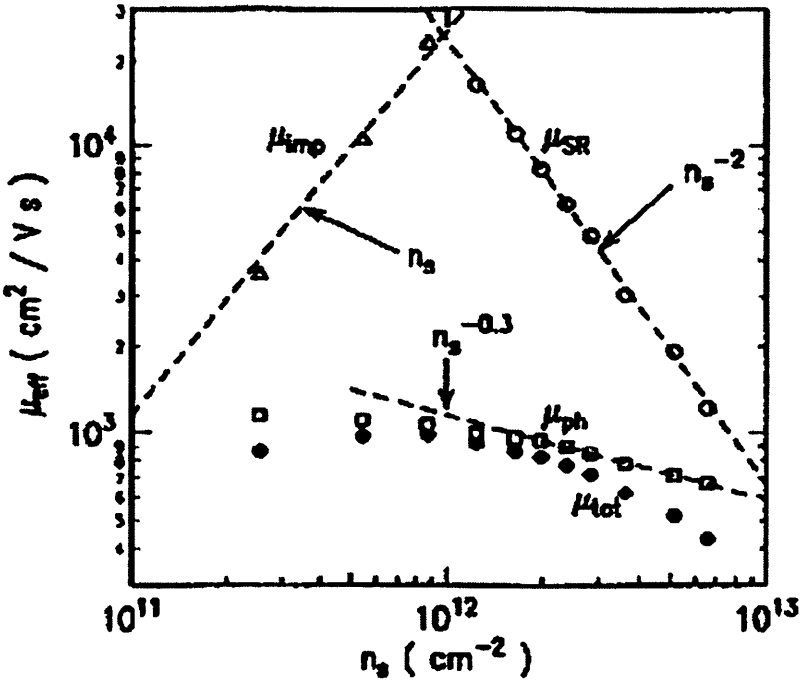


Fig. 4.3. This plot adapted from [4.8] shows that phonon mobility is proportional to the inversion charge density to the power –0.3. Since $N_{INV} \propto E_{EFF}$ when the device is in the strong inversion regime, we conclude $\mu_{PH} \propto E_{EFF}^{-0.3}$.

The theory of electron-phonon interactions in semiconductor quantum wells [4.9] predicts that the phonon scattering rates are proportional to the electron density of states. The density of states per unit energy rate in a 2-D quantum well is given by [4.9]

$$N(E) = \frac{nm^*}{2\pi\hbar^2 L}, \tag{4.4}$$

where n is a quantum number, m^* is the electron effective mass and L is the width of the rectangular well. In MOS applications, the inversion electrons are confined to a triangular well (Fig. 3.1). Thus we replace L with the effective size x_C of the well given by equation (3.8). x_C is proportional to $E_{EFF}^{-1/3}$, the phonon scattering rate is then proportional to $E_{EFF}^{1/3}$, and

$$\mu_{PH} \propto E_{EFF}^{-1/3}, \tag{4.5}$$

in agreement with the simulation results in Figure 4.3.

4.2.3 Interface scattering

Phonon scattering alone cannot account for the experimentally observed universal mobility at the Si-SiO_2 interface. At high vertical electric field, the electron mobility starts to drop more rapidly then the $E_{EFF}^{-1/3}$-dependence would predict. To account for this fact it has been proposed that surface roughness accounts for the increased rate of scattering at high E_{EFF} [4.10]. Asperities at the Si surface create spatially varying potential along the interface that scatters conduction electrons. The scattering probability, according to the model, is proportional to the second power of the root-mean-square of the roughness Δ, and to the second power of the roughness correlation length L_C. To explain the experimental data, the value of Δ = 4.3Å and L_C = 15Å should be used [4.7].

The model also predicts that interface mobility is proportional to E_{EFF}^{-2}, and provides a good fit to the experimental data.

One major concern with the validity of this model is that it predicts a very strong mobility dependence on Δ. It is therefore puzzling why electron mobility appears to be rather insensitive to the precise oxide growth conditions as well as to the smoothness of the original Si wafer. As a matter of fact, the micro-roughness of Si wafers has been substantially reduced over the past decades. Currently, the micro-roughness measured by atomic force microscopy is below 2Å [1.1]. At the same time, electron mobility has not improved over the past three decades.

Electron mobility does not seem to be degraded by a moderate increase in surface roughness either. In one such experiment [4.12], the Si surface roughness was intentionally increased (by 30-hour DI water rinse) from 5Å to 20Å. Electron mobility however remained unchanged (Fig. 4.4). We therefore believe that, at reasonably smooth interfaces, a different mechanism is responsible for mobility degradation at high vertical electric field.

4.3 Electron wavefunction penetration into the gate dielectric

The alternative interface scattering model that we develop in the following chapter is based on carrier wavefunction penetration. When quantum mechanical simulations for the electron distribution in the inversion layer are performed, a zero boundary condition for the electron wavefunction is assumed at the Si-dielectric interface. This is equivalent to assuming that the gate dielectric has an infinite barrier height. In this work, we relaxed the infinite-barrier assumption by modifying the quantum

mechanical simulator [2.14], to determine the amount of carrier wavefunction in the gate dielectric. Figure 4.5 illustrates that a small amount of electron wavefunction penetrates into SiO_2.

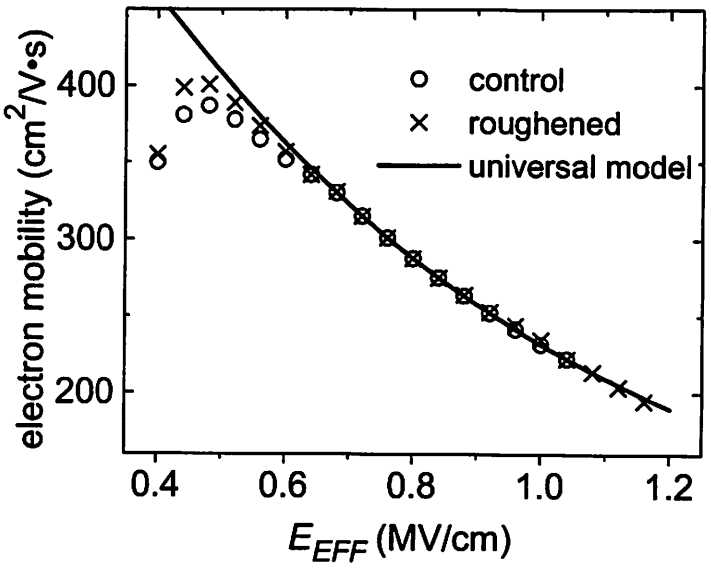


Fig. 4.4. Experimental mobility measurements adapted from [4.12] show that electron mobility is not affected by surface roughness.

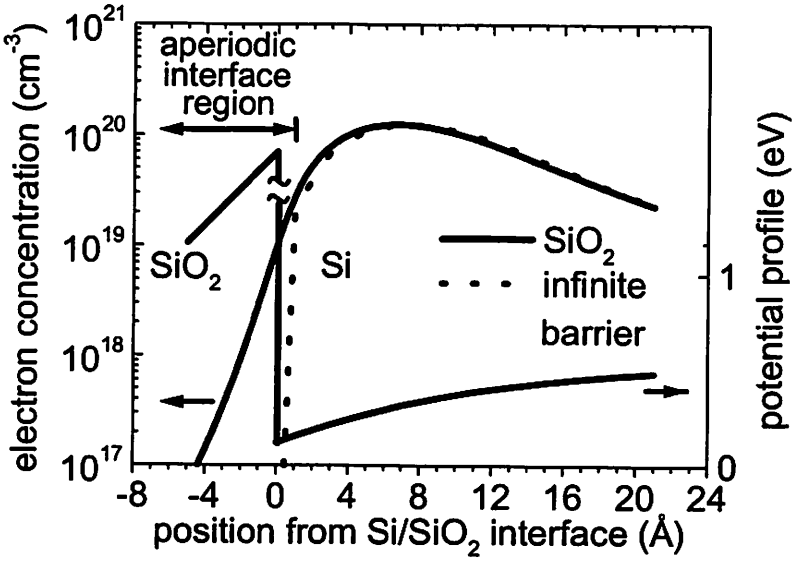


Fig 4.5. Electron wavefunction penetration into gate dielectric.

We quantify the amount of wavefunction penetration by calculating the relative fraction of the electron density outside of the periodic Si lattice.

$$f = \frac{\int_{-\infty}^{boundary} \rho \cdot dx}{\int_{-\infty}^{+\infty} \rho \cdot dx}. \tag{4.6}$$

As one might already expect based on the discussion in section 3.3.2, the fraction f depends only on a single parameter E_{EFF} (Fig. 4.6). Interestingly, this turns out to be a simple linear dependence

$$f = s \cdot E_{EFF}, \tag{4.7}$$

where the slope parameter s depends of the barrier height and effective electron mass of the dielectric. Since high-κ dielectrics have lower barrier height (and typically lower effective mass) than SiO_2 (Table 4.1), greater wavefunction penetration is expected for these dielectrics (Fig. 4.7).

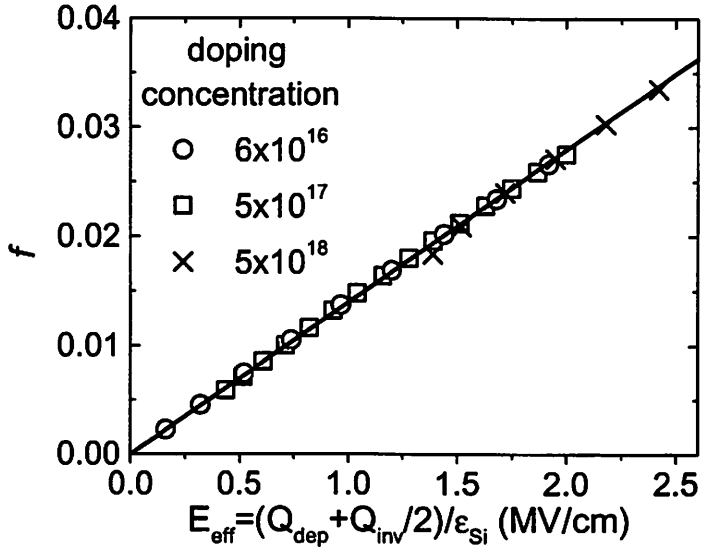


Fig. 4.6. Wavefunction penetration fraction f is independent of the channel doping concentration and is determined exclusively by E_{EFF}.

TABLE 4.1. Dielectric barrier heights and effective carrier masses for electrons and holes in SiO_2, Si_3N_4, and HfO_2.

Dielectric	ϕ_{Be}, eV	ϕ_{Bh}, eV	m_e	m_h
SiO_2	3.1	4.1	0.4	0.32
Si_3N_4	2.1	1.9	0.5	0.41
HfO_2	1.1	3.4	0.2	Unknown

The effects of wave function penetration on the electron distribution in the inversion layer include an increase of the inversion charge density and the shift of the charged centroid closer to the dielectric interface [4.14] (Fig. 4.5). In this chapter however, we are going to focus exclusively on the effects of wavefunction penetration on the electron channel mobility.

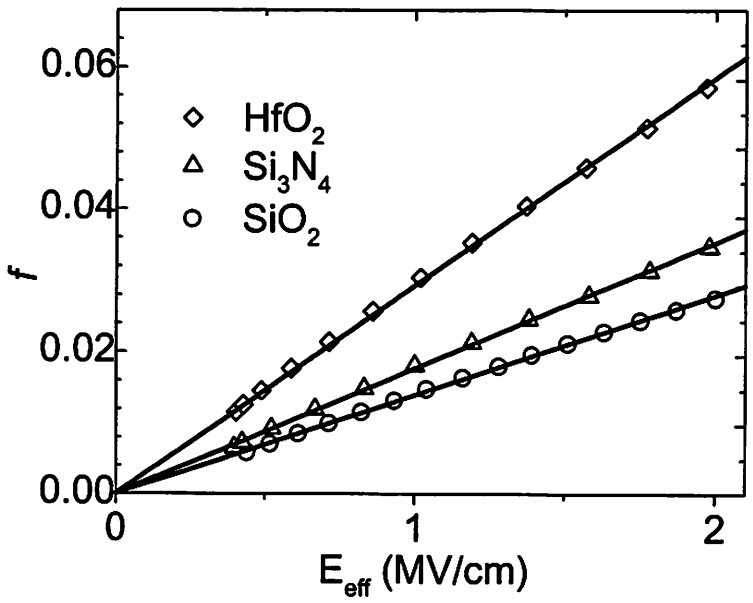


Fig. 4.7. Fraction f of electron wavefunction in the various dielectrics is a linear function of E_{EFF}. There is a larger wavefunction penetration into high-κ dielectric.

4.4 New Interface Scattering model

Most of the inversion electron wavefunction propagates within the periodic Si lattice, where under ideal circumstances (at zero temperature) electrons do not experience any scattering. A small fraction of the electron wavefunction propagates within the amorphous dielectric where the scattering rates are much higher than in a periodic crystalline medium. Even the periodicity of the top monolayer of Si atoms is disturbed by the amorphous material above it; therefore we include the top 1Å of Si in the "aperiodic interface region" in this study.

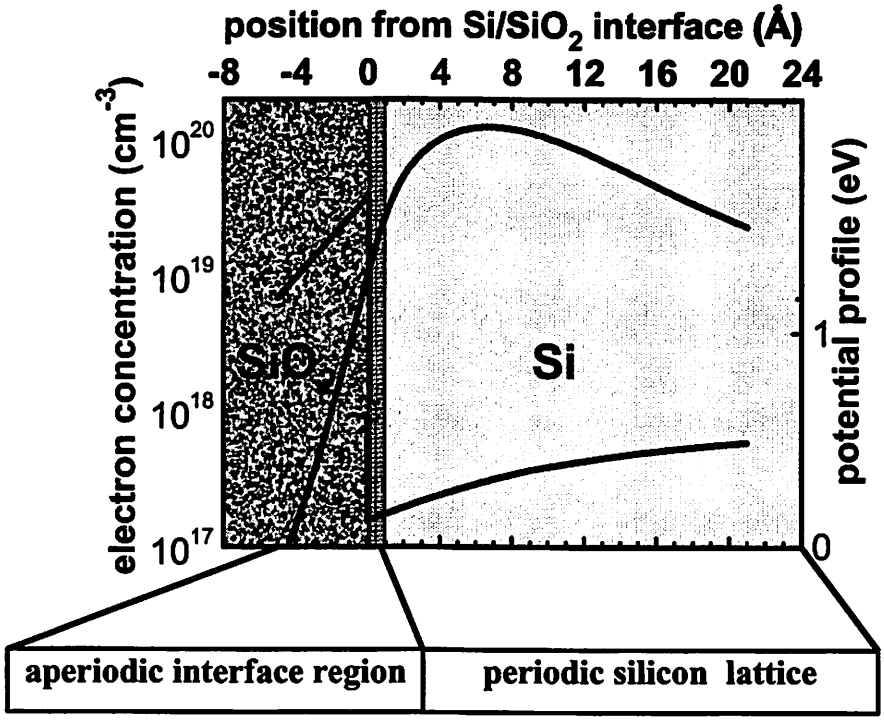


Fig. 4.8. Small fraction of electron wavefunction in the aperiodic interface region is susceptible to a very high probability of scattering, and this can substantially increase the total scattering rate for the inversion layer electrons.

Scattering probability P in the aperiodic region is proportional to the square of the matrix element for the corresponding Hamiltonian H. The precise form of the Hamiltonian is unknown. However it can be replaced by a constant (similar to deformation potential for phonon scattering) within the aperiodic region; and H is zero outside of this region, so that the integration extends only over the region to the left of the boundary between the periodic and aperiodic regions.

$$P \propto \left\langle \Psi_i \mid H \mid \Psi_f \right\rangle^2 \propto \left(\int_{-\infty}^{boundary} \Psi(x)\Psi^*(x)dx \right)^2 = f^2 = s^2 \cdot E_{eff}^2 \qquad (4.8)$$

The interface mobility is inversely proportional to this scattering probability, and is therefore proportional to E_{EFF}^{-2}, as required by the universal mobility model.

$$\mu_{INT} = \mu_{OX} f^2 = \mu_{OX} s^{-2} E_{EFF}^{-2}. \qquad (4.9)$$

Here μ_{OX} is a fitting parameter related to the Hamiltonian H. If we combine μ_{INT} with phonon mobility μ_{PH} according to the Mathiessen's rule (4.2) we find that our model correspond precisely to the experimentally established universal mobility model [4.15] (Fig. 4.9). The value $\mu_{OX} = 0.12$ cm^2/(V·s) was used, which is a factor of 170 less than the calculated mobility in crystalline SiO_2 [4.16]. (For comparison, that ratio of mobilities in single-crystalline and amorphous Si is 1000).

4.5 Universal mobility model for high-κ dielectrics

The proposed model can be extended to high-κ dielectrics. Since the amount of the wavefunction penetration for these materials is higher than for SiO_2 (Fig. 4.7), lower mobility can be expected. A calculated mobility curve for HfO_2 is also shown in Figure 4.7. (The same value of μ_{OX} is assumed for HfO_2 as for SiO_2.) The mobility for HfO_2 gate

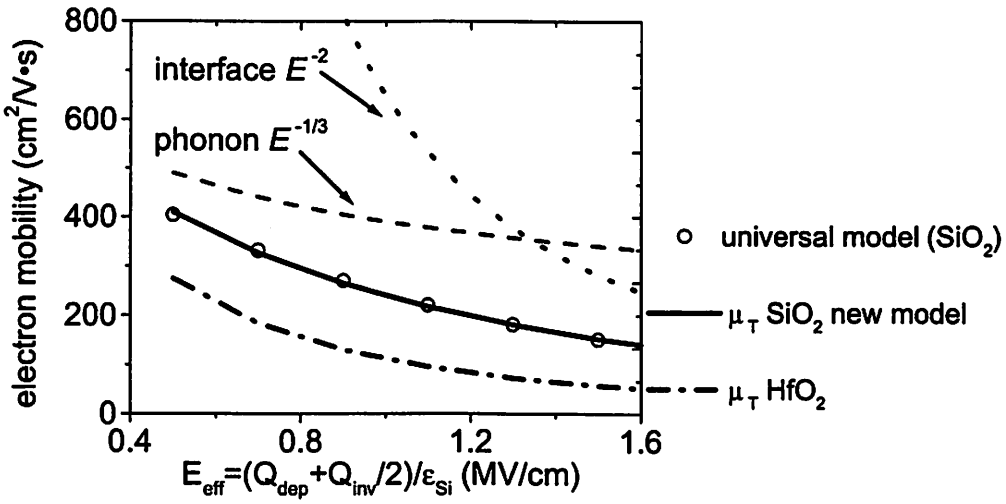


Fig. 4.9. Proposed interface scattering mechanism matches the universal mobility model for SiO_2 and allows to develop universal mobility models for other gate dielectrics.

dielectric is about half of that for SiO_2. This mobility results are similar to the reported experimental results [4.17].

As a matter of fact, the comparison between the model and the experimental data for high-κ dielectrics is difficult, because there is usually an interfacial layer of unknown composition formed between the substrate and high-κ dielectric. Nevertheless, we try to use the available experimental data to support the proposed model in the next section.

4.6 Model application

4.6.1 Application to gate oxynitrides

Gate dielectrics for various NMOSFET samples described in this section were formed in a 4-step process: First, the interfacial layer was formed by rapid thermal NO

oxidation, then Si_3N_4 was deposited by an RTCVD process. This was followed by a 900°C NH_3 anneal, and N_2O oxidation. (Details are provided in Table 4.2). The mobility was extracted by split-CV method.

TABLE 4.2. Gate dielectric formation conditions.

Wafer	NO pressure (Torr)	RTCVD deposition	900°C NH_3 anneal	N_2O anneal temperature (°C)
A1	100			900
A2	100			850
A3	100			800
B1	1			900
B2	1			850
B3	1			800

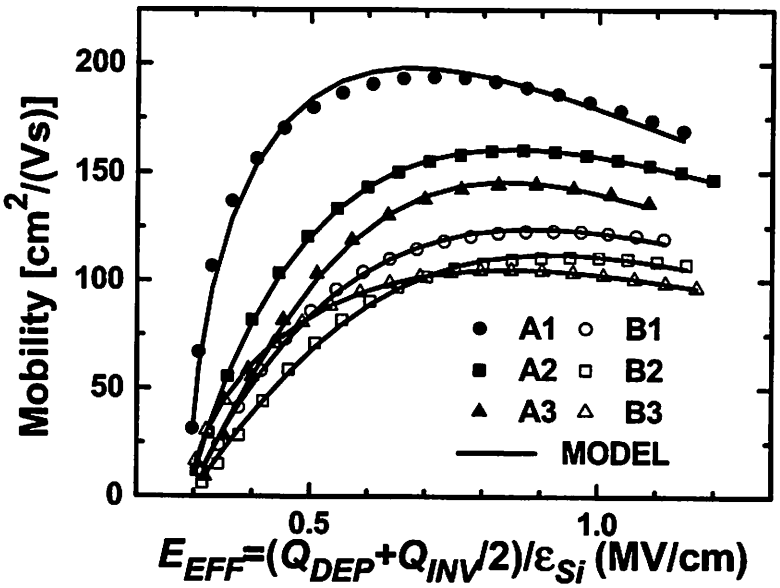


Fig. 4.10. The quantitative mobility model (4.2) matches a wide range of experimental data.

The mobility model (4.1) was used fit the experimental data (Fig. 4.10). The Coulombic component μ_{COUL} and the interface component were modeled according to (4.3) and (4.9) respectively. The phonon component is described by

$$\mu_{PH} = 407 \cdot E_{EFF}^{-0.3} \frac{\text{cm}^2}{\text{V} \cdot \text{s}}, \tag{4.10}$$

when E_{EFF} is expressed in MV/cm. The mobility model provides an excellent fit to measured data (Fig 4.10). Furthermore, both Coulombic and interface components of mobility are strongly correlated to the nitrogen content in the dielectric film (Fig. 4.11). (The relative nitrogen content was determined by XPS analysis for three of the dielectric films.) It is well established that higher nitrogen content leads to increased densities of the defects and interface states in the dielectric, leading to an increase in Coulombic scattering.

It has been shown that higher nitrogen content improves the smoothness of the Si-dielectric interface [4.18]. Thus the experimental correlation in this study contradicts the surface roughness model. Instead, it supports the wavefunction model described in the previous section. As nitrogen content in the dielectric increases, the dielectric barrier height is reduced, leading to greater wavefunction penetration and consequently lower mobility.

Figure 4.12 illustrates how the nitrogen content in the film and the interface mobility are affected by the processing conditions. First, we observe that higher pressure NO oxidation results in higher μ_{INT}. This agrees with previous studies that reported a more 'oxide-like' interface (less nitrogen) for the oxynitride layers grown at higher NO pressure [4.19]. Second, the higher temperature of post-deposition N_2O anneal improves

μ_{INT}, as it enhances the growth of a new interfacial layer with smaller nitrogen content than that of the original NO-grown interface [4.19, 4.20].

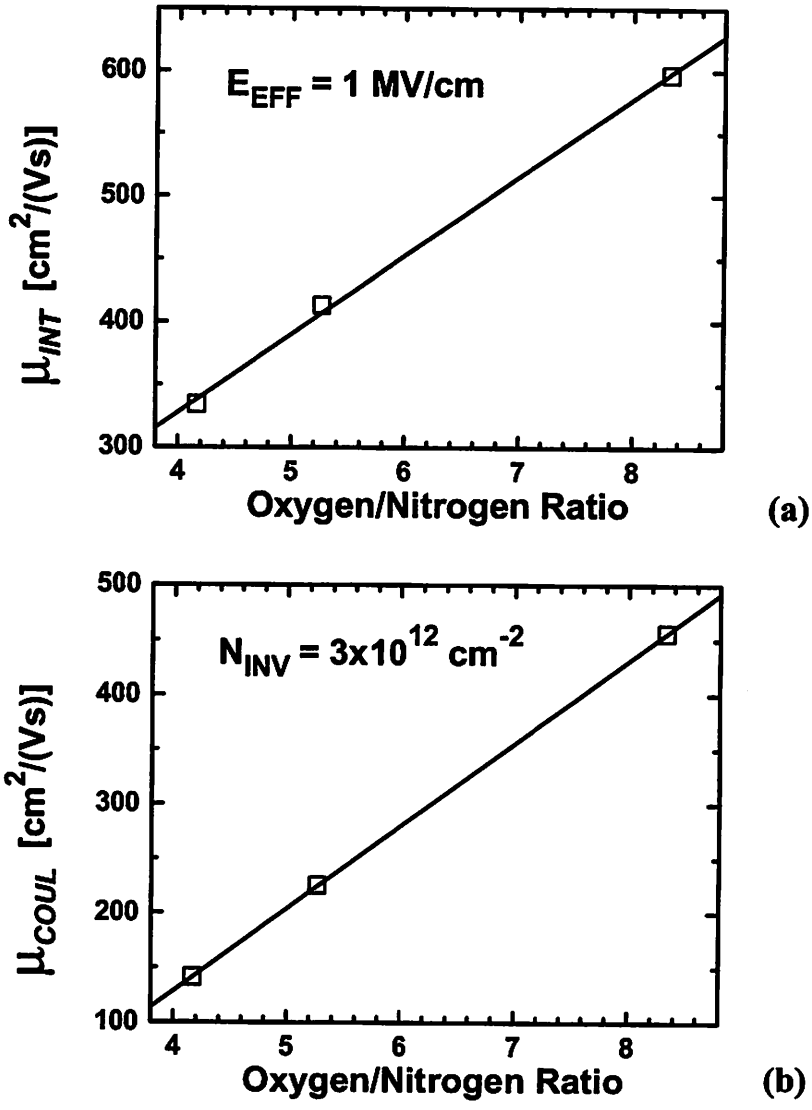


Fig. 4.11. Both Coulombic (a) and interface (b) mobility depend on the nitrogen content of the dielectric. Lower nitrogen concentration results in higher mobility.

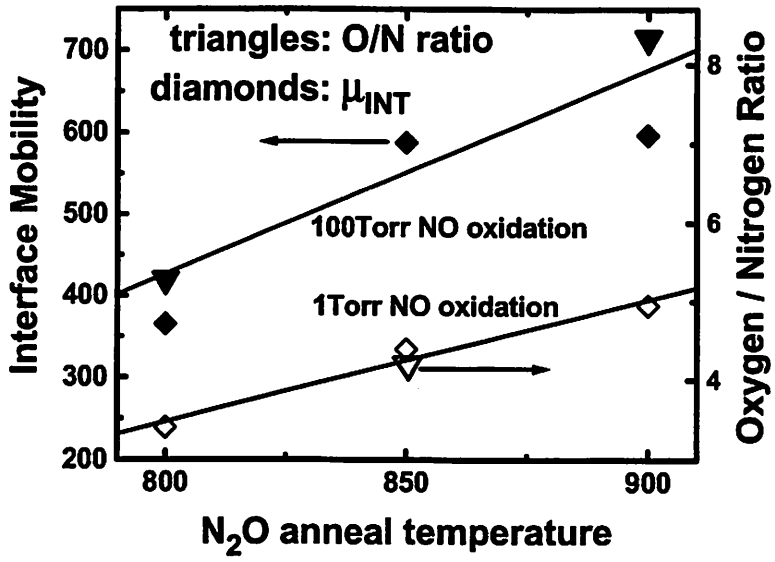


Fig. 4.12. Higher temperature N_2O anneal reduces nitrogen concentration near the interface, increases dielectric barrier height and improves interface mobility.

4.6.2 Application to epitaxial $SrTiO_3$ gate dielectric.

According to the new interface scattering model the degradation in carrier mobility comes about due to the high scattering rates in the amorphous gate dielectrics. Therefore if an epitaxial single-crystalline gate dielectric is introduced instead, a substantial improvement in the channel mobility at high effective electric field should be expected.

Transistors with epitaxial $SrTiO_3$ gate dielectric have been recently reported [4.21]. Our model can successfully match the mobility for these devices (Fig. 4.13). It is interesting to note that while mobility for $SrTiO_3$ at low E_{EFF} is strongly degraded, the mobility at high E_{EFF} rivals that of SiO_2. This can be explained based on the electron

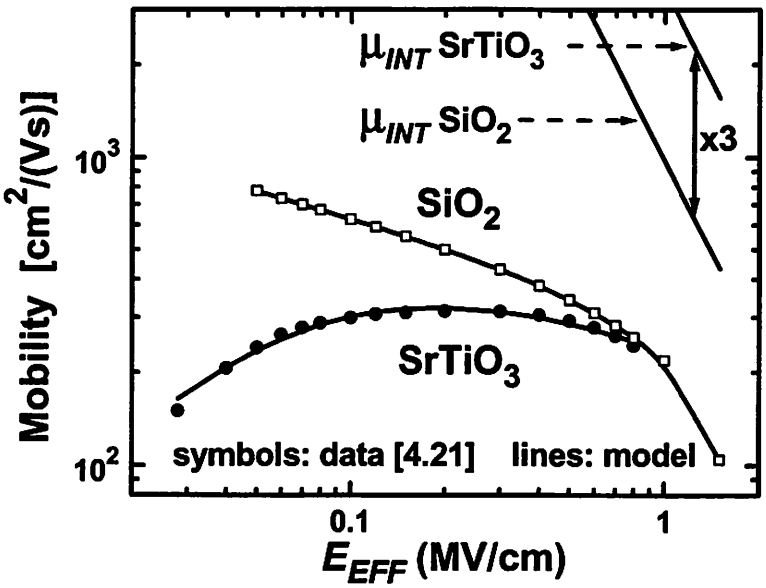


Fig. 4.13. Epitaxial gate dielectrics [4.21] can have higher mobility at large E_{EFF}, since the scattering by the aperiodic atoms in the amorphous dielectric is effectively eliminated.

wavefunction penetration scattering model. Even though $SrTiO_3$ has a low barrier height (less than 1eV), and consequently a large amount of wavefunction penetration, the scattering rate for this single-crystalline material is much smaller than that for amorphous SiO_2. As a result, μ_{INT} is a factor of 3 higher for $SrTiO_3$ than for SiO_2 (Fig. 4.13).

4.7 Conclusion

The proposed quantitative mobility model provides a very good fit to the experimental data over the entire range of gate biases, and can be applied to any dielectric stack (high-κ with oxynitride interface, as well as high-κ directly on Si). Furthermore, our experimental results indicate that electron wavefunction penetration rather then

surface roughness is the mechanism responsible for mobility degradation at high gate bias. In order to ensure high carrier mobility at the Si – gate dielectric interface the following three approaches should be considered:

1. An interfacial SiO_2 layer can be introduced between the channel and the high-κ dielectric. The disadvantage of this approach is that the presence of the interfacial SiO_2 layer can reduce the scaling limits of the gate stack equivalent-oxide thickness, as discussed in Chapter 3.

2. If aggressive dielectric scaling requires that high-κ be deposited directly on Si, then a dielectric with larger barrier height and larger effective electron mass is needed to suppress the wavefunction penetration. This requirement does not add any extra restrictions on the dielectric choice, since large barrier heights and effective masses are also required to suppress the tunneling current (equations 3.17 and 3.18).

3. Single crystalline dielectrics could be epitaxially grown on silicon. Because lower scattering rates are expected for crystalline materials, the channel carrier mobility should improve. Formation of epitaxial dielectrics on Si still remains relatively unexplored. In addition high quality epitaxial films would be needed, since the grain boundaries and other defects can assist the tunneling leakage current through the gate dielectric.

4.8 References

4.1 S. Takagi, A. Toriumi, M. Iwase, H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: Part II – effects of surface orientation," *IEEE TED* Vol. 41, No. 12, pp. 2363-2368, Dec. 1994.

4.2 A.G. Sabnis and J.T. Clemens, "Characterization of Electron Velocity in the Inverted <100> Si Surface," *Tech. Dig. International Electron Devices Meeing,* pp. 18-21, 1979.

4.3 E. D. Siggia, P. C. Kwok, "Properties of electrons in semiconductors inversion layers with many occupied electric subbands. I. Screening and impurity scattering," *Phys. Rev. B*, Vol. 2, No. 4, pp. 1024-1036, 15 Aug. 1970.

4.4 T. H. Ning, C. T. Sah, "Theory of scattering of electrons in a nondegenerate-semiconductor-surface inversion layer by surface-oxide charges," *Phys. Rev. B*, Vol. 6, No. 12, pp. 4605-4613, 15 Dec. 1972.

4.5 F. Gamiz, J. A. Lopez-Villanueva, J. A. Jimenez-Tejada, I. Melchor, and A. Palma," A comprehensive model for Coulomb scattering in inversion layers," *J. Appl. Phys.*, Vol. 75, No. 2, pp. 924-934, 15 Jan, 1994.

4.6 T. Matsuoka, S. Taguchi, Q. D. M. Khosru, K. Taniguchi, C. Hamaguchi, "Degradation of inversion layer electron mobility due to interface traps in metal-oxide-semiconductor transistors," *J. Appl. Phys.*, Vol. 78, No. 5, pp. 3252-934, 1 Sep. 1994.

4.7 T. Ando, A. B. Fowler, F. Stern, "Electronic properties of two dimensional systems," *Reviews of Morden Phys.*, Vol. 54, No. 2, pp. 437-672, April 1982.

4.8 M. V. Fischetti, S. E. Laux, "Monte Carlo study of electron transport in silicon inversion layers," *Tech. Dig. International Electron Devices Meeing,* pp. 721-724, 1992.

4.9 B. K. Ridley, "The electron –phonon interaction in quasi-two-dimensional semiconductor quantum-well structures," *J. Phys. C: Solid State Phys.*, Vol. 15, pp. 5899-5917, 1982.

4.10 A. Hartstein, T. H. Ning, A. B. Fowler, "Electron scattering in silicon inversion layers by oxide and surface roughness," *Surf. Sci.* Vol. 58, No. 1, pp. 178-181. Aug. 1976.

4.11 The International Technology Roadmap for Semiconductors, Semiconductor Industry Association, San Jose, CA, 2001.

4.12 F. Assaderaghi, D. Sinitsky, J. Bokor, P. K. Ko, H. Gaw, C. Hu, "High-field transport of inversion-layer electrons and holes including velocity overshoot," *IEEE Transactions on Electron Devices*, Vol. 44, No. 4, pp. 664-71, April 1997.

4.13 available at http://www-device.eecs.berkeley.edu/qmcv.html

4.14 I. Polishchuk, C. Hu, "Electron wavefunction penetration into gate dielectric and interface scattering-an alternative to surface roughness scattering model," *Symp. on VLSI Technology*, pp. 51-52, Kyoto, Japan, June 2001.

4.15 M.S. Liang, J.Y. Choi, P.K. Ko, and C. Hu, "Inversion-Layer Capacitance and Mobility of Very Thin Gate-Oxide MOSFET's," *IEEE Trans. Electron Devices*, ED-33, No. 3, pp. 409-413, Mar. 1986.

4.16 L. Scozzoli, S. Reggiani, M. Rudan, "Homogeneous transport in silicon dioxide using the spherical-harmonics expansion of the BTE," *IEICE Transactions on Electronics*, Vol. E83-C, No. 8, pp. 1183-1188, Aug. 2000.

4.17 C. Hobbs, *et al.*, "80 nm Poly-Si Gate CMOS with HfO_2 Gate Dielectric," *Tech. Dig. International Electron Devices Meeing,* pp. 651-654, 2001

4.18 R. Hedge, P. Tobin, K. Reid, B. Maiti, S. Ajuria, "Growth and surface chemistry of oxynitride gate dielectric using nitric oxide," *Appl. Phys. Lett.*, Vol. 66, No. 21, pp. 2882-1884, 22 May 1995.

4.19 L. Gosset et al., "High resolution depth profiling in silicon oxynitride films using narrow nuclear reaction resonances," *Nuc. Inst. and Meth. of Phys. Res. B,* Vol. 136, pp. 521-527, March 1998.

4.20 V. Misra, W. K. Henson, E. M. Vogel, G. A. Hames, P. K. McLarty, J. R. Hauser, J. J. Wortman, "Elecrical properties of composite gate oxides formed by rapid thermal processing,", *IEEE Trans. Electron Devices*, Vol. 43, No. 4, pp. 636-646, Apr. 1996.

4.21 R. A. McKee, F. J. Walker, M. F. Chisholm, "Physical structure and inversion charge at a semiconductor interface with a crystalline oxide," *Science*, Vol. 293 pp. 468-471, 2001.

Chapter 5 CMOS Gate Technology Based on Metal Interdiffusion

5.1 Background and Motivation

There are two important factors that have, over the years, influenced the choice of the gate electrode material for MOS devices. The first factor is the work function (or Fermi energy) of the electrode material. It plays an important role in determining the electrical characteristics of the devices. The second factor is the thermal stability of the electrode material on the gate dielectric. This factor influences the processing conditions under which the devices are fabricated.

The first MOS transistors reported by Kahng and Atalla in 1960 actually had a metal gate electrode [5.1]. Since a non-self-aligned process was used, and thermal stability was not an issue. However, as the channel length of the MOS transistors began to shrink, the alignment of the gate to the source and drain regions became more critical. A self-aligned process required that gate electrode be stable on SiO_2 during the high-temperature source/drain formation step. Consequently, polycrystalline silicon became the gate material of choice for decades to come [5.2]. Another remarkable innovation was the introduction of CMOS technology [5.3], whereby a considerable reduction in the

circuit power consumption was achieved by combining NMOS and PMOS transistors. As CMOS technology developed and supply voltages were reduced, it became necessary to use different work function gate electrodes for NMOS and PMOS devices. Hence, the longevity of polysilicon as a gate electrode material can also be attributed to the fact that by doping polysilicon, its work function can be changed so that desired threshold voltages for NMOS and PMOS transistor can be achieved simultaneously. While polysilicon gate electrodes have been hugely successful over the past 3 decades, they might not be able to meet the increasing stringent requirements that CMOS scaling is going to impose over the next decade.

From the device performance perspective, polysilicon-gated transistors suffer from what is know as polysilicon depletion effect, illustrated in Figure 5.1. When an MOS structure is biased into strong inversion, a strong electric field is developed across the gate dielectric. This field pushes the mobile carriers in the gate electrode away from the gate dielectric interface. This effectively increases the capacitance-equivalent thickness of the gate dielectric, and as a result has a detrimental effect on device performance. From the processing perspective, polysilicon has been shown to be unstable on many of the advanced gate dielectrics, such as Zr_2O [5.4] and Ta_2O_5 [5.5].

Since metals have a very high electron concentration, metal gates can essentially eliminate the polysilicon depletion effect. Metal gates are also expected to be more stable than polysilicon on many of the high-κ gate dielectrics.

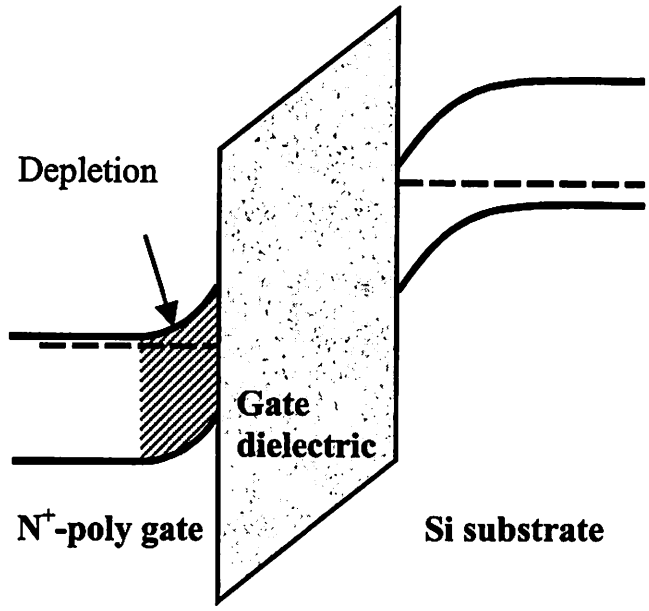


Fig. 5.1. MOS band diagram under strong inversion bias. A layer depleted of mobile carriers is formed in the polysilicon gate. This depletion layer increases the total capacitance-equivalent oxide thickness by a few angstroms.

5.2 Work function requirements

In order to built high-performance CMOS circuits, it is required that both NMOS and PMOS devices have low and symmetric threshold voltages (V_T's). This, in turn, requires that the gate electrodes have two different work functions. For bulk NMOS and PMOS transistors the desired work function values should be around 4.1 eV and 5.2 eV respectively [5.6]. When the gate material is silicon (or another semiconductor), such dual work functions can be achieved by doping the gate with either donor or acceptor atoms. Changing the work function of a metal is much harder; therefore it is likely that two metals have to be used in order to implement a dual-work-function metal gate

technology. Therefore new methods and techniques for making gate electrodes have to be explored.

Furthermore, for future CMOS technologies the choice of the gate work functions is not necessarily going to be limited to the two values mentioned above. The choice of the work function is likely to be dictated by the transistor structure, as well as by circuit architecture. In order to scale CMOS transistors down to 15 nm channel length, alternative device structures such as Fin-FETs [5.7] and ultra-thin-body transistors [5.8] have been proposed. These advanced transistors will require work function values of around 4.4 eV and 4.9 eV [5.9].

In order to reduce the power consumption in high performance CMOS circuits, the use of multiple threshold voltages has been proposed [5.10]. The use of low threshold voltage transistors can be limited only to the circuit's critical path, while transistors with high threshold voltage can be used elsewhere, thus reducing the total power consumption without compromising circuit performance. The multiple threshold voltage technique is also a promising approach for system-on-chip applications.

5.3 Metals' work functions and chemical properties

By definition, the work function of a metal is the minimum amount of energy required to remove an electron from the surface of that metal. Chemical properties of a metal are strongly dependent on its ability to donate electrons to form electronic bonds. Thus, it is natural to expect that both the work function of a metal and its chemical properties, to a large extent, are determined by the valance electron binding energy.

Consequently, one would expect a strong correlation between the work function of a metal and its chemical properties

Chemical properties of an atom are often expressed by the element's electronegativity, or its power to attract electrons to itself. Electronegativities of selected metals [5.11] are plotted against their work functions in Figure 5.2. The observed correlation between the metal work functions and their chemical properties poses certain challenges for the successful integration of metal gate technology in CMOS process.

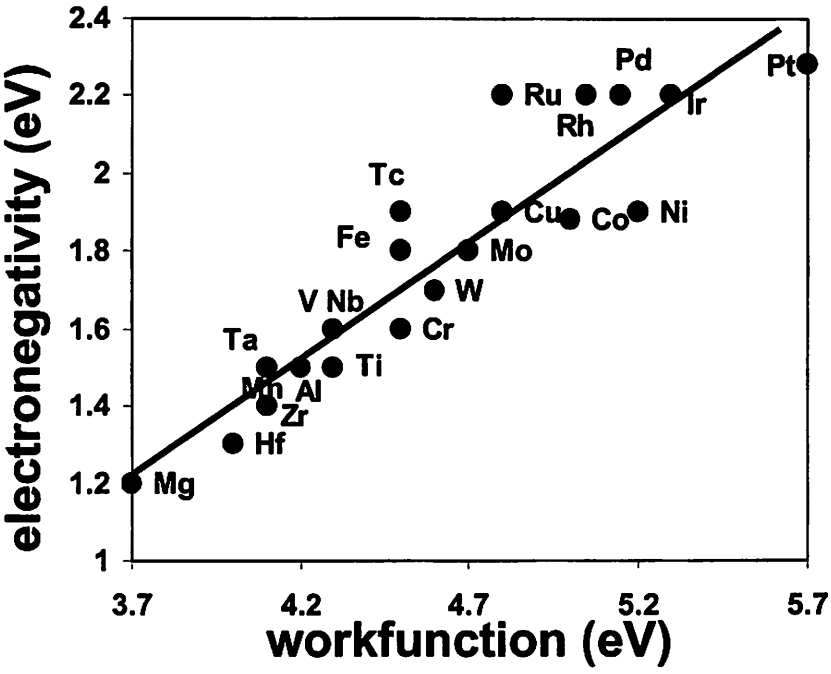


Fig. 5.2. Chemical properties of metal are strongly linked to their work function. High work function metals (PMOS candidates) are inert; low work function metals (NMOS candidates) are reactive.

We first note that the high-work-function metals (the PMOS candidates) are typically chemically inert. As a consequence, these materials will be difficult to etch during the gate definition process. In addition, the high-work-function metals, such as Pt for example, are known to have adhesion problems to SiO_2. The most viable metals for bulk PMOSFETs are therefore the ones that have high work function (above 4.9 eV) and a moderate electronegativity. Nickel, cobalt, and iridium appear to be the most promising metals.

Integration of a low-work-function metal is even more challenging. These metal have low electronegativity values, and are therefore very reactive. In order to avoid the reaction a low-work-function metal and gate SiO_2 dielectric during the high-temperature processing steps certain modification of the standard CMOS process may be required. First, a replacement gate process can be used. In such a process, the gate stack is formed after the high temperature steps required for source drain activation have been completed; and a possible reaction between the metal gate electrode and the gate dielectric is thus averted. Second, an alternative gate dielectric can be used to replace SiO_2. If hafnium oxide (HfO_2) is used, for example, then any metal with the electronegatity equal or higher than that of Hf is expected to be stable. These metals include most of the low-work-function metals, such as, Hf, Zr, Ta, Ti; and all of the high-work-function metals.

5.4 Metal Interdiffusion Process

While much of the recent metal gate research has focused on identifying the best candidate metals for NMOS and PMOS transistors [5.12-5.15], the task of integrating two metals into a CMOS process flow remains largely overlooked. When two different

gate electrode materials are used for NMOS and PMOS transistors new integration issues arise. To illustrate these issues, we shall first consider the simplest conceivable way of fabricating dual work function CMOS transistor, point out the shortcomings of this approach, and finally propose an alterative and more advantages way of dual work function metal gate transistors.

A straightforward way to implement dual work function metal gate CMOS [5.16] is demonstrated in Fig. 5.3. After blanket deposition, the first metal is removed from either the PMOS or NMOS side. Then a second metal with a different work function is deposited. Now two different metals determine the threshold voltages of the NMOS and PMOS devises. This simple fabrication process, unfortunately, entails exposing the gate dielectric to the etchant, when the first metal is removed. The etchant will very likely cause undesirable thinning of the gate dielectric. This in turn will result in the increase in the gate leakage current and potential dielectric reliability problems.

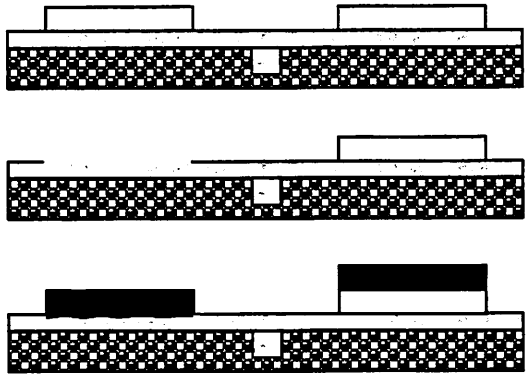
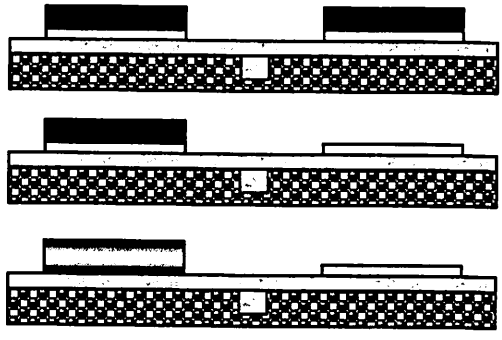


Fig. 5.3. A simple approach to dual metal gate integration involves etching one of the metals from the dielectric surface. This is likely to jeopardize the integrity of the gate dielectric.

For a successful dual metal gate process, it is important to be able to find a way to get the second metal down to the dielectric interface without first stripping the first metal. This goal can be accomplished by means of metal interdiffusion. The proposed metal interdiffusion gate (MIG) process is illustrated in Fig. 5.4. First a thin layer of one of the metals is deposited over the entire wafer. For sake of discussion, let us assume this first metal is the one with the low work function. Then the second (high-work-function) metal is deposited over the entire wafer. Next, the high work function metal is selectively removed from the NMOS side while the PMOS side is protected by photoresist. Since the low-work-function metal is the only metal remaining on top of the NMOS dielectric it will clearly determine the NMOSFET's threshold voltage (V_T). The two remaining metals on the PMOS side are subsequently allowed to interdiffuse.

In general, the two metals will form an alloy that has an intermediate work function. By choosing a suitable thickness combination for the metal layers one can control the



a) deposit metal 1, then metal 2

b) remove metal 2 from NMOS side

c) anneal to diffuse metal 2 to the dielectric interface

Fig. 5.4. The improved approach to dual metal gate integration relies on metal interdiffusion. No metal has to be etched from the dielectric surface, thus the gate dielectric is always protected.

alloy composition and thus the gate work function. This approach would allow for a continuous tuning of transistor threshold voltages independent of substrate doping concentration. This can be especially important for ultra-thin body MOS transistors [1.4], where the substrate doping is not an effective way to control V_T.

In certain cases, one of which we explore in the following sections, the top-layer metal has a propensity to segregate at the dielectric interface. Consequently, this metal will solely determine the work function of the p-MOS gate electrode.

5.5 Ni-Ti MIG-FET

5.5.1 Capacitor Fabrication

To demonstrate the proposed dual work function technology we chose Ti as a low-work-function metal (for NMOS) and Ni as a high-work-function metal (for PMOS) to fabricate MOS capacitors. In order to precisely extract the work functions of the metal gate electrodes (see section 5.5.2), SiO_2 capacitors with 4 different gate dielectric thicknesses on each of the wafers were fabricated in the following manner: First, a 55-nm layer of SiO_2 was thermally grown on (100) Si wafer. Then the wafer was immersed into HF halfway (Fig. 5.5) and the bottom half of the wafer is etched. After the etch, the wafer is rinsed in DI water and dried. Finally, the wafer was rotated by 90° and once again immersed into HF halfway (Fig. 5.5), etched, and dried. As long as the two etch times are not equal SiO_2 layers of 4 different thicknesses are obtained on the different quadrants of the wafer.

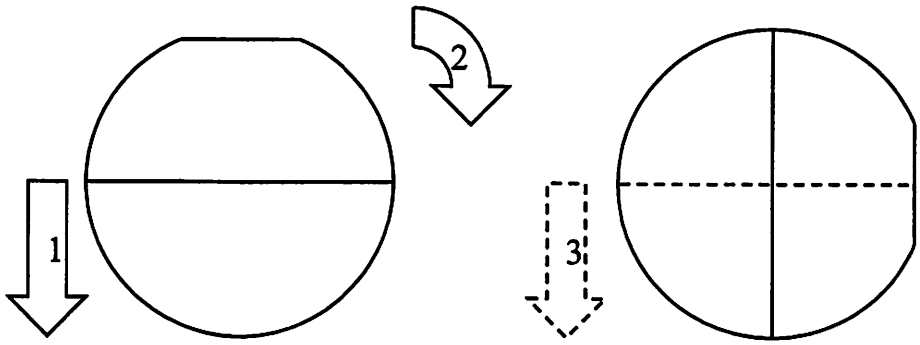


Fig. 5.5. A 2-step etch-back process is used to obtain four different oxide thicknesses on a single wafer. The wafer is rotated by 90° between the first and the second etch.

The fabrication sequence for the Ti-Ni interdiffusion gate electrode is as follows: First, an ultra thin (100 Å) layer of Ti is sputter-deposited on top of the SiO_2 gate dielectric followed by 200 Å of Ni. The sample is then annealed in forming gas ($N_2 + H_2$) at 400°C for 30 minutes. In addition, 2 control samples with pure Ti (100 Å) and pure Ni (100 Å) gates were fabricated. In an attempt to prevent the oxidation of Ni and Ti in the gate stacks, a 200 Å cap-layer of TiN was sputter-deposited on all samples.

5.5.2 Work Function extraction

The work functions of the metal gates can be determined by performing capacitance versus gate voltage (CV) measurements and extracting the flat-band voltage (V_{FB}) of an MOS capacitor [5.18]. A large shift in the flat-band voltage of a PMOS capacitor with Ti/Ni gate resulting from metal interdiffusion at 400°C can be observed in Figure 5.6. The shift in the flat-band voltage can be attributed in part to the change in the

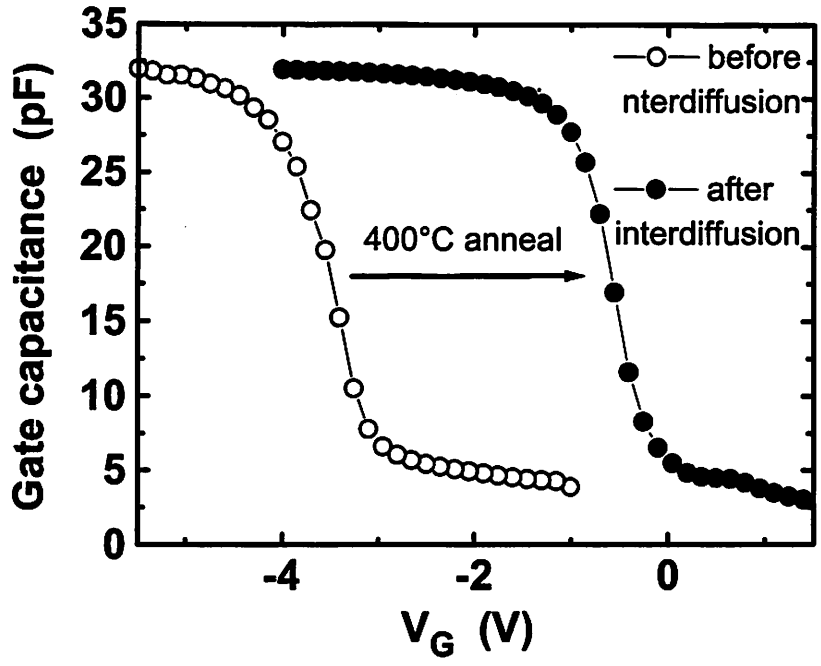


Fig. 5.6. Capacitance versus voltage characteristic of a fabricated MOS capacitor with Ti/Ni gate before and after interdiffusion anneal. (t_{ox}=35 nm.)

the work function of the gate electrode and in part to the change in the oxide fixed charge Q_f, since V_{FB} is given by

$$V_{FB} = \phi_M - \phi_S - Q_f\, t_{ox}/\varepsilon_{ox} \qquad (1)$$

Here ϕ_M is the work function of the metal gate, the work function of the p-type silicon substrate ϕ_S is equal to 4.9 V in our devices and ε_{ox} and t_{ox} are the permittivity and thickness of the gate oxide respectively.

The contribution from the fixed charge Q_f can be eliminated by measuring V_{FB} for multiple oxide thicknesses on the same wafer; the intercept of the V_{FB} versus t_{ox} plot gives the $\phi_M - \phi_S$ value. Figure 5.7 shows that before interdiffusion the flat-band voltage

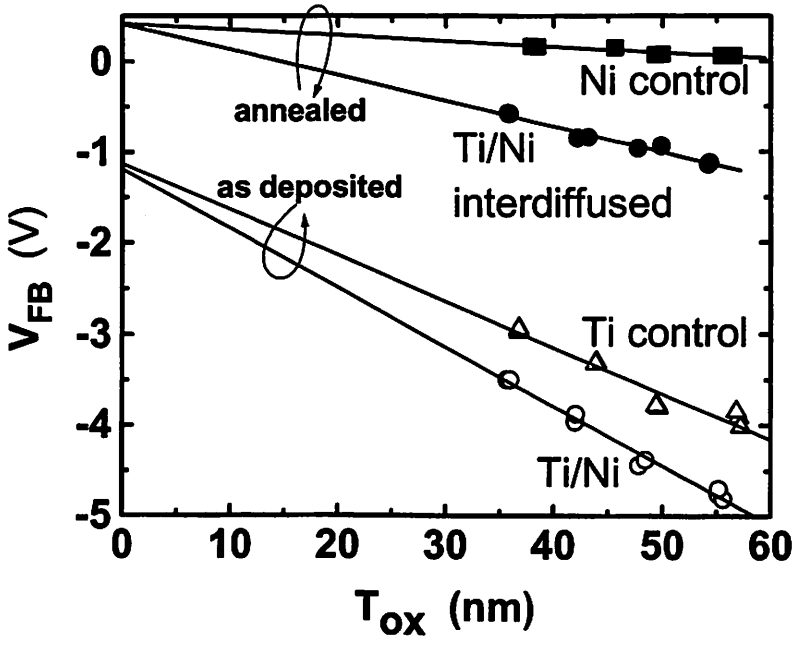


Fig. 5.7. Before annealing, the work function of the Ti/Ni gate is the same as that of the Ti-control sample; after annealing, it is the same as that of the Ni-control sample.

of Ti/Ni electrode corresponds to that of the Ti-gate control sample, since Ti is still on the bottom of the metal gate stack (Fig 5.4.a). After the 400°C anneal the flat-band voltage of the Ti/Ni gate corresponds precisely to that of the Ni-gate control sample annealed under the same conditions. This observation indicates that as the result of the 400°C anneal, Ni has diffused to the SiO_2 interface and now determines the gate work function (Fig 5.4.c). The measured Ti work function is 3.9 eV and corresponds to the silicon conduction band (Fig. 5.8), making Ti a good choice for the bulk NMOSFET gate electrode. Ni has a work function of 5.3 eV (close to the silicon valence band), and can be used to make PMOSFETs with a low threshold voltage.

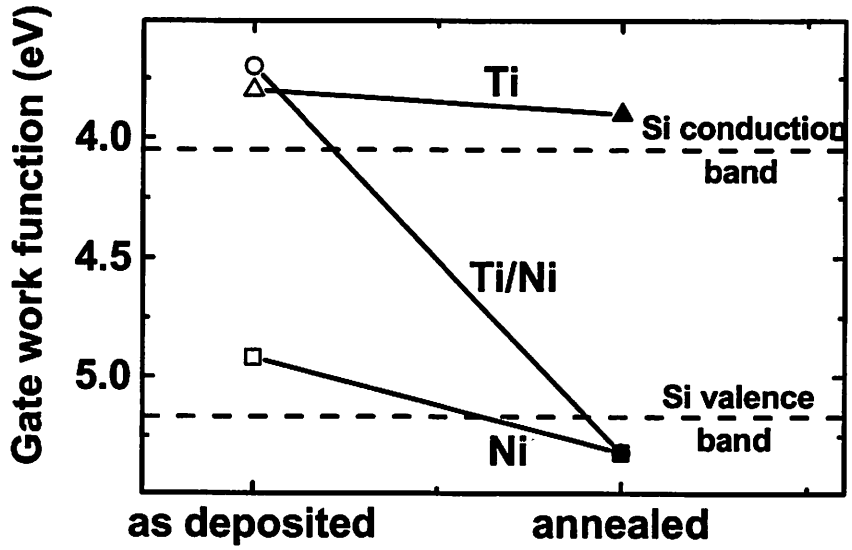


Fig. 5.8. Work function of Ti lies near Si conduction band and is appropriate for bulk-NMOS gate application. The work function of Ti/Ni gate after interdiffusion lies near Si valence band and is appropriate for bulk-PMOS gate application.

5.5.3 X-ray Photoelectron spectroscopy analysis

We further confirm the interdiffusion of titanium and nickel by X-ray photoelectron spectroscopy (XPS) analysis. The XPS analysis was performed at the Materials Analysis Group of Accurel Systems using a Surface Science Instruments SSX-1000 spectrometer. The depth profile shown in Figure 5.9 was gathered by alternating electron spectroscopy measurement and ion sputtering. Note that the x-axis scale of Figure 5.9 is obtained by converting the sputter time to sputter depth using the known sputter rate for SiO_2, and therefore does not represent the actual depth of the metal gate stack. The depth profile for the annealed Ti/Ni sample shown indicates that Ni, which

was originally sandwiched between the Ti and TiN layers, has diffused to both the top and bottom interfaces of the gate electrode. The diffusion of Ni through the TiN layer can be explained by the fact that our TiN film was nitrogen-deficient and thus did not serve as a good diffusion barrier. A more important conclusion is that Ni has segregated to the dielectric interface. This explains why the Ti/Ni gate electrode has a high work function after the anneal. The detailed photoelectron spectrum for Ni is shown in Figure 5.10. The three lines represent the spectra measured at the top, in the middle, and at the bottom of the gate electrode. Ni at the top of the gate electrode is partially oxidized due to the exposure to air. Only elemental Ni is present in the middle of the gate electrode and at the SiO_2 interface. The absence of nickel oxide or nickel silicide at the dielectric interface indicates that Ni electrode is stable on SiO_2.

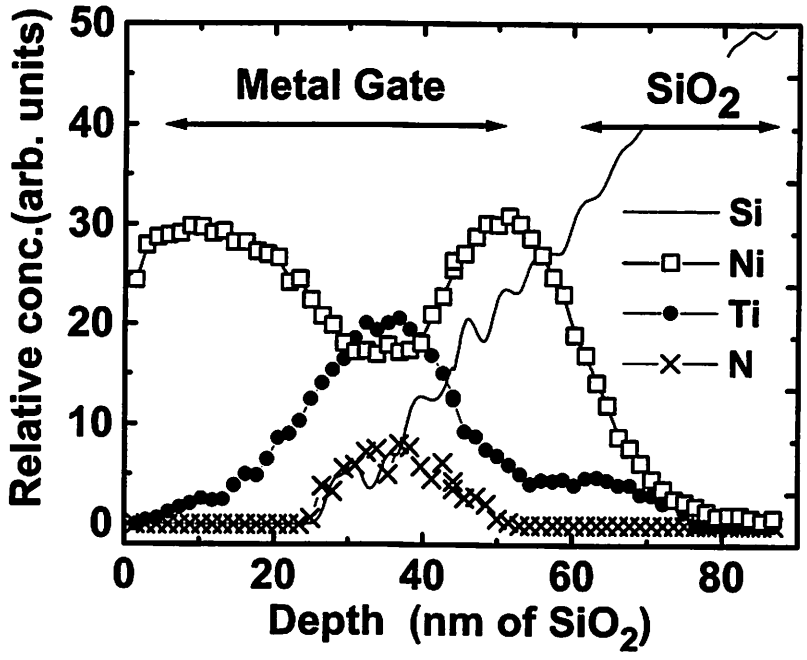


Fig. 5.9. XPS depth profile for Ti/Ni gate electrode after the 400°C, 30 min. interdiffusion anneal. A large concentration of Ni is present at the SiO_2 interface, thus Ni determines the gate work function.

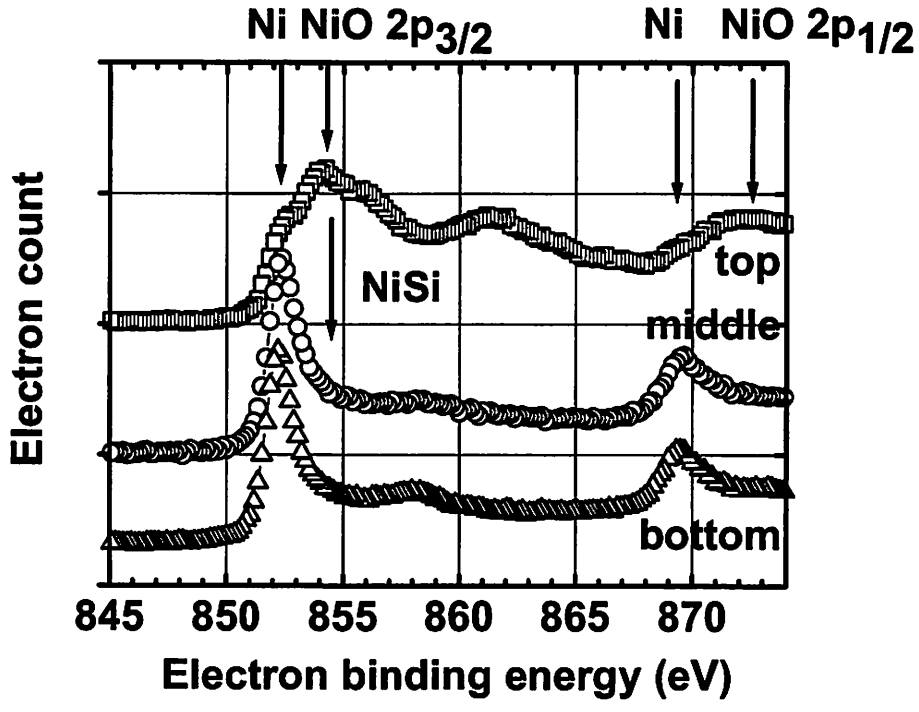


Fig. 5.10. The photoelectron spectrum taken at the bottom of the MIG electrode shows that only elemental Ni is present at the dielectric interface. This indicates that Ni is thermally stable on SiO_2.

5.5.4 Transistor fabrication and characterization

While Ni is thermally stable on SiO_2, as discussed in the previous section, Ti is known to react with SiO_2 at temperatures above 400°C. Therefore, to demonstrate the feasibility and the advantages of the metal interdiffusion gate (MIG) CMOS technology, we used a simple non-self-aligned gate-last process. NMOS and PMOS transistors in this experiment were fabricated on separate wafers in order to minimize the number of masking steps, while the gate interdiffusion process described in section 5.4 is designed to work for a single CMOS wafer as well. The fabrication process flow is shown in figure 5.11.

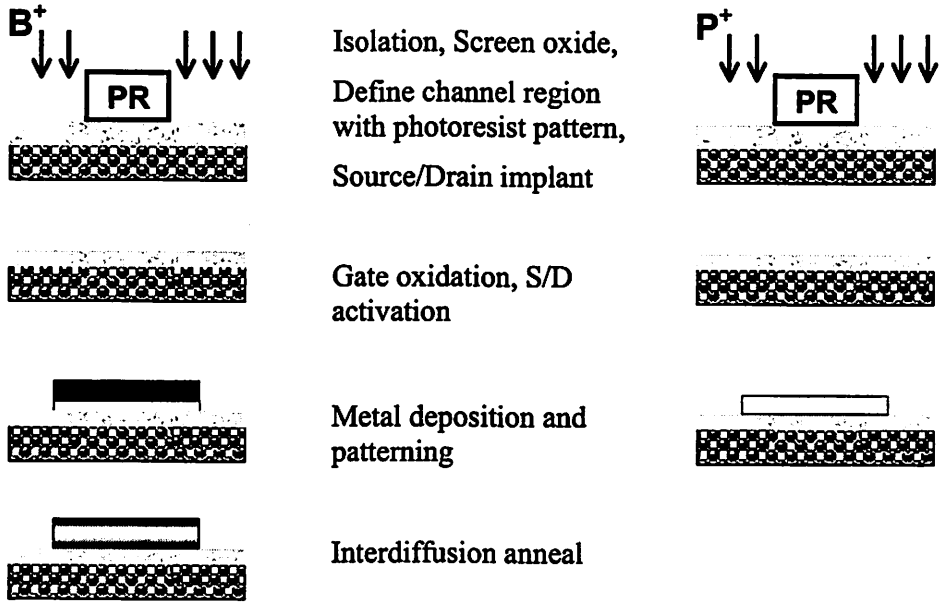


Fig. 5.11. Illustration of the fabrication process flow for MIGFETs.

The thicknesses of the Ti and Ni layers used in this experiment were 80 Å and 200 Å respectively. A 15-minute 400°C interdiffusion anneal was performed in the forming gas ambient. A low V_T of about 0.5V is achieved for NMOS transistors with Ti gate, as can be seen from the CV characteristic (Fig. 5.12.a). As deposited, the PMOS Ti/Ni gate electrode has an effective gate work function corresponding to that of Ti (since Ti is the bottom layer), resulting in a high threshold voltage (Fig. 5.12.c). However, after the interdiffusion anneal, Ni segregates to the gate dielectric interface, and from that point on determines the gate work function, resulting in a low V_T of about -0.5V for the PMOS transistors.

In addition to the experimental CV data (symbols), Figures 5.12.a and 5.12.b show the results of quantum mechanical simulations of CV characteristics for CMOS

transistors with both metal (non-depleting) gates (solid lines) and polysilicon gates (dashed lines). Close match between the experimental data and the simulation for the non-depleting gates in the deep inversion regime indicates that metal gates successfully eliminate polysilicon depletion for both NMOS and PMOS devices. This results in a significant (10%) increase in the inversion capacitance as compared to polysilicon gated devices with 25 Å gate oxide. The advantage will be even more significant for the devices with thinner gate dielectrics.

The fabricated long-channel NMOS and PMOS transistors have well-behaved I_D-V_D characteristics (Fig. 5.13). The NMOSFETs show excellent turn-off characteristics; however, the PMOSFETs show large off-state leakage. This leakage is attributed to tunneling between the drain and the gate in the large overlap region, and can be avoided by using a self-aligned fabrication process.

Since metal penetration through the gate dielectric during the interdiffusion anneal is a potential issue for the proposed MIG technology, we have investigated the P-MOS channel hole mobility. The fact that it closely matches the universal mobility model [5.19] (Fig. 5.14) indicates that there is no metal penetration into the channel region.

The shortest channel transistors fabricated in this non-self-aligned process have a channel length of 0.8 μm and show good drive current (Fig. 5.15).

Since metal penetration through the gate dielectric during the interdiffusion anneal is a potential issue for the proposed MIG technology, we have investigated the P-MOS channel hole mobility. The fact that it closely matches the universal mobility model [5.19] (Fig. 5.14) indicates that there is no metal penetration into the channel region.

The shortest channel transistors fabricated in this non-self-aligned process have a channel length of 0.8 μm and show good drive current (Fig. 5.15).

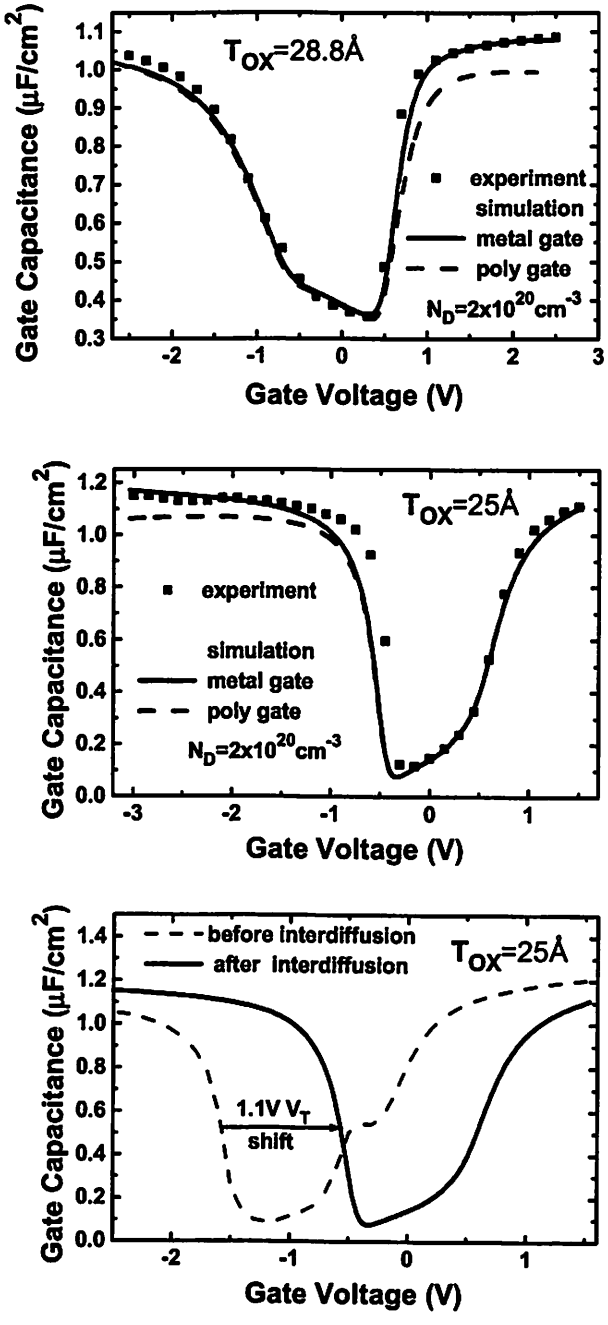


Fig. 5.12. Capacitance vs. voltage characteristics for **(a)** n-MOSFET and **(b)** p-MOSFET show that the transistors have low and symmetric V_T's. Inversion capacitance is well matched by the quantum mechanical simulations. Metal gate increases the inversion capacitance by 10% through elimination of the polysilicon depletion. **(c)** p-MOSFET *C-V* characteristics before and after metal interdiffusion anneal. A V_T shift of 1.1V is observed corresponding to the change in the effective gate work function.

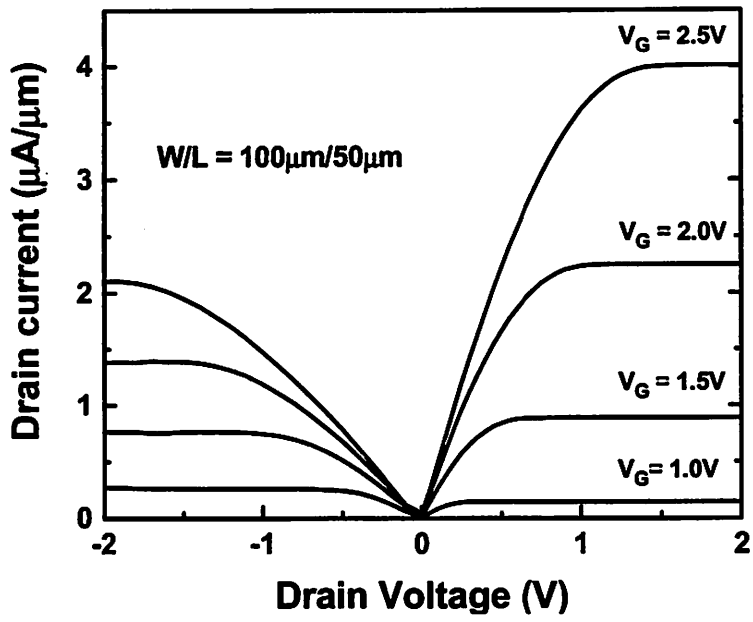


Fig. 5.13. I_D-V_D characteristics of the long channel NFET and PFET.

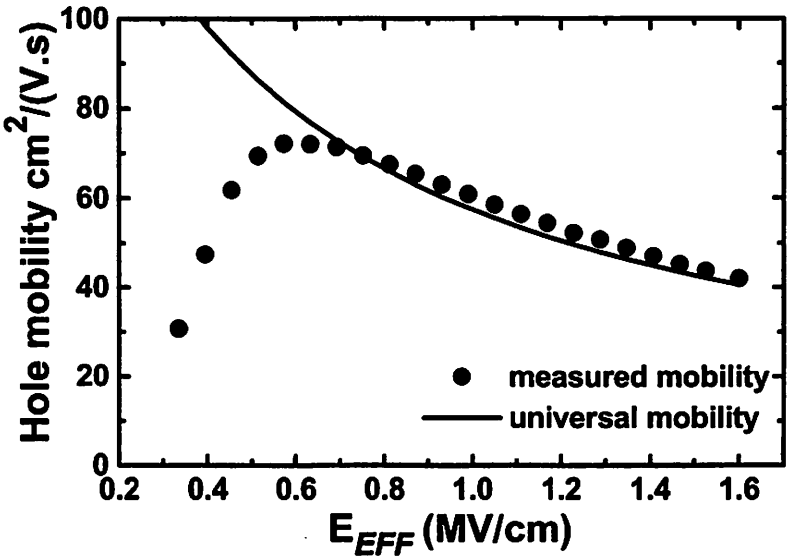


Fig. 5.14. Measured PMOSFET hole mobility is in excellent agreement with the universal mobility model indicating no significant metal penetration into the channel region.

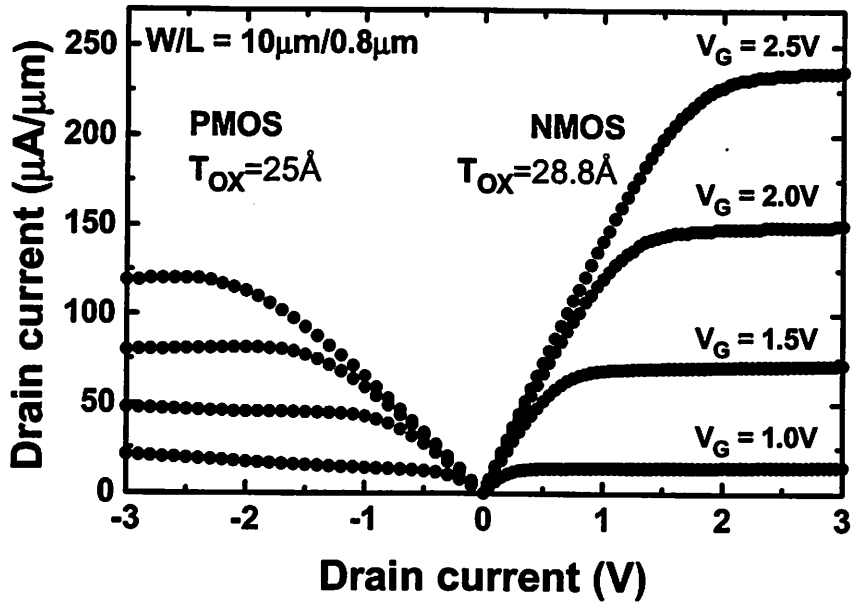


Fig. 5.15. CMOS transistors with L_{GATE}= 0.8μm show high drive current indicating the suitability of the MIG process for short channel devises.

5.5.5 Examination of physical mechanisms

In this section we shall try to examine the possible mechanisms responsible for the successful implementation of the proposed metal interdiffusion process for the Ni-Ti system. There are clearly two independent phenomena that contribute to the final outcome. One is the high diffusivity of Ni in Ti at low temperatures; the other is Ni segregation to the dielectric interface.

It has long been established that Ni diffuses in Ti exceptionally fast [5.20, 5.21]. In fact, Ni in Ti diffuses faster than any other metals reported in those studies (Co, Fe, Mn, Cr, Nb, Mo, Sc, V). The diffusion coefficient of Ni in Ti has been reported in the 900°C to 1700°C temperature range (Fig. 5.16). Two diffusion mechanisms (vacancy-assisted and interstitial) can be identified. The extrapolation of the interstitial diffusion

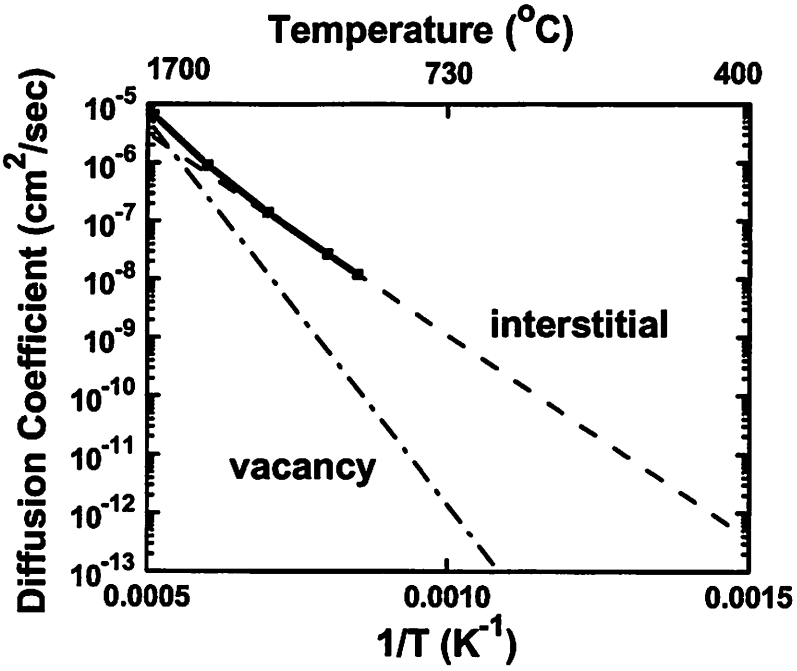


Fig. 5.16. Diffusion coefficient of Ni in Ti. Experimental data [5.21] is shown with solid line, extrapolation to low temperatures is shown with dashed line.

(low temperature mechanism) down to 400°C results in a diffusion constant value of 4×10^{-13} cm^{-2}/sec. This corresponds to diffusion length of 1800Å during a 15 minute anneal, a distance large enough to ensure Ni diffusion all the way to the dielectric interface in our experiments.

The second issues that remains to be examined is the segregation of Ni to the SiO_2 interface. Even though the particular reason for segregation might be difficult to establish, we should at least address the question of whether this segregation is always going to occur in the Ni-Ti system. To answer this question the following experiment was performed. Capacitors with Ni-Ti stack gate electrodes were fabricated as described earlier in this chapter. However in this case, Ni was deposited on the bottom. In addition,

the nickel layer was thin (60 Å), and the titanium layer was thick (1000 Å). If the nickel's propensity to segregate to the SiO_2 interface were absolute, Ni would still remain at the dielectric interface even after the 400°C interdiffusion anneal.

The result of the work function extraction before and after the anneal are shown in Figure 5.17. The work function of the gate electrode after the deposition is 4.7 eV, similar to what was observed for the Ni-control gate (after deposition). After the anneal, however, the work function becomes 4.0 eV similar to the Ti work function. This indicates that the gate electrode near the dielectric interface is primarily composed of Ti.

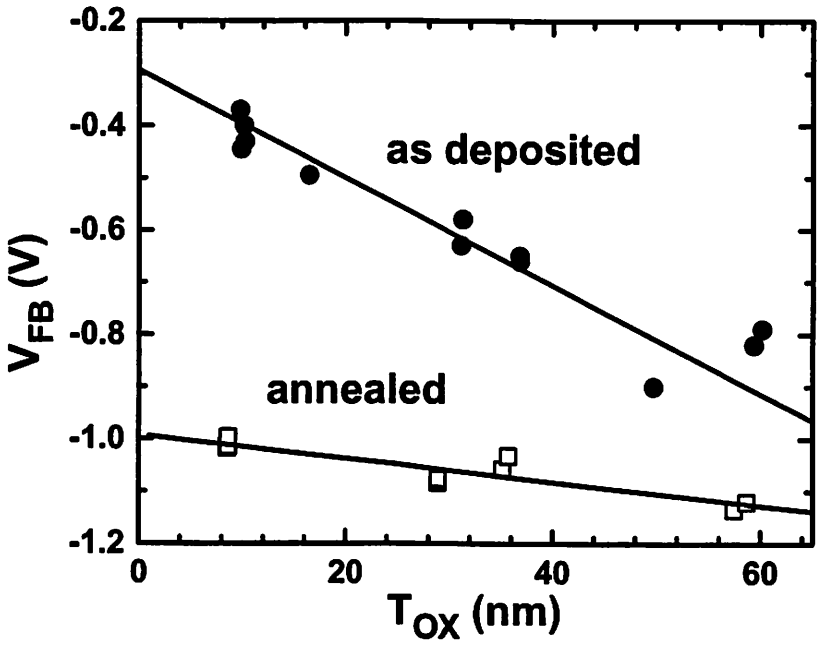


Fig. 5.17. As deposited the Ni/Ti gate electrode has a high work function (flat band voltage) since Ni is at the dielectric interface. After a 5 min 400°C anneal Ni diffuses away from the interface and the gate work function drops by 0.7 eV.

Evidently, there are two competing mechanisms that determine the concentration of Ni at the interface. On one hand, Ni has a propensity to segregate to SiO_2 interface (perhaps due the fact that it has lower surface energy), and that is what determines the final work function in the original experiment. On the other hand, lower bulk energy of an intermetallic compound drives the diffusion of Ni away from the SiO_2 interface (This factor is dominant in the second experiment.) The outcome clearly depends on whether we have a Ni-rich or a Ti-rich alloy.

To understand the difference between Ni-rich or Ti-rich alloys we shall examine the Ni-Ti phase diagram [5.22] (Fig 5.18). The large triangular region on the right hand side of the diagram represents the solid solution of Ti and Ni. This is most likely the phase obtained during the interdiffusion anneal in the original experiment. As the mixture

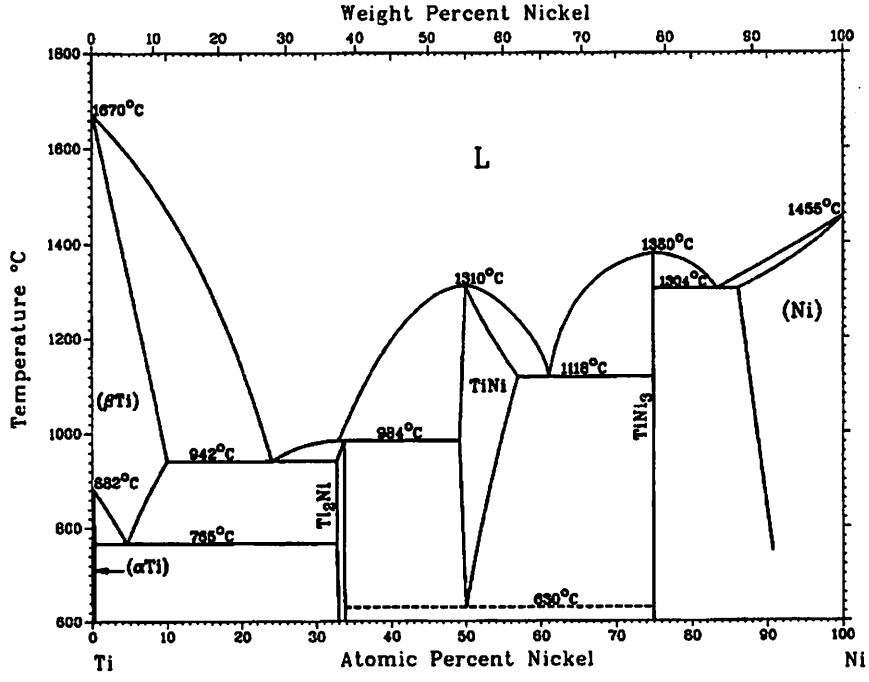


Fig. 5.18. Ti-Ni phase diagram [5.22].

is cooled down the solid solubility limit of Ti drops. Excess Ti segregates out of the solid solution to form the $TiNi_3$ phase (presumably away from the dielectric interface). Therefore, the layer adjacent to the SiO_2 interface becomes composed almost exclusively of Ni; and this explains the observed high work function.

In the second experiment we clearly deal with a Ti-rich alloy that is composed of two phases Ti and Ti_2Ni. The solid solvability limit of Ni in Ti does not change much when the sample is cooled, so that the composition of these two phases remains the same. Either or both of these phases are eventually present at the dielectric interface; and they determine the observed low work function of 4eV.

5.6 Mo-Nb metal system

While the Ni-Ti system has remarkable diffusion properties, titanium, as mentioned in section 5.5.4, is not stable on SiO_2 at high temperatures. It is therefore, important to explore other metals for potential applications (especially NMOS applications) with SiO_2 gate dielectric. According to Figure 5.2 Nb has a larger electronegativity than Ti does, and consequently is expected to be more stable on SiO_2. Molybdenum has been reported to have a work function compatible with PMOS requirements and excellent thermal stability [5.23]. In addition, both Mo and Nb can have a wide range of work functions depending on their crystalline orientations (Table 5.1). Potentially, the crystalline orientation can be controlled by choosing an appropriate deposition method and conditions.

In our Mo-Nb interdiffusion experiment the metal gate capacitors were fabricated by first evaporating an 80Å layer of Nb followed by a 500Å of Mo. The samples were

then subjected to a series of anneals at 400°C, 600°C, 800°C, and 900°C. The measured work functions are reported in Figure 5.19. The original gate work function of 4.65eV corresponds closely to the value reported for (112) Nb [5.11] (Table 5.1). Upon the series of anneals the measured work function increases gradually. This gradual increase is indicative of the Mo-Nb intermixing, rather than a change in the crystalline orientation of Nb which would likely result in a more sudden work function change. Furthermore, the final gate work function of 4.95eV matches that of (110) Mo. Since formation of (110) Mo on SiO_2 following a high temperature anneal has been confirmed [5.23], we can conclude with a high degree of certainty that Mo has diffused to the dielectric interface.

These results indicate that the proposed metal interdiffusion gate process can be implemented in a variety of metal systems.

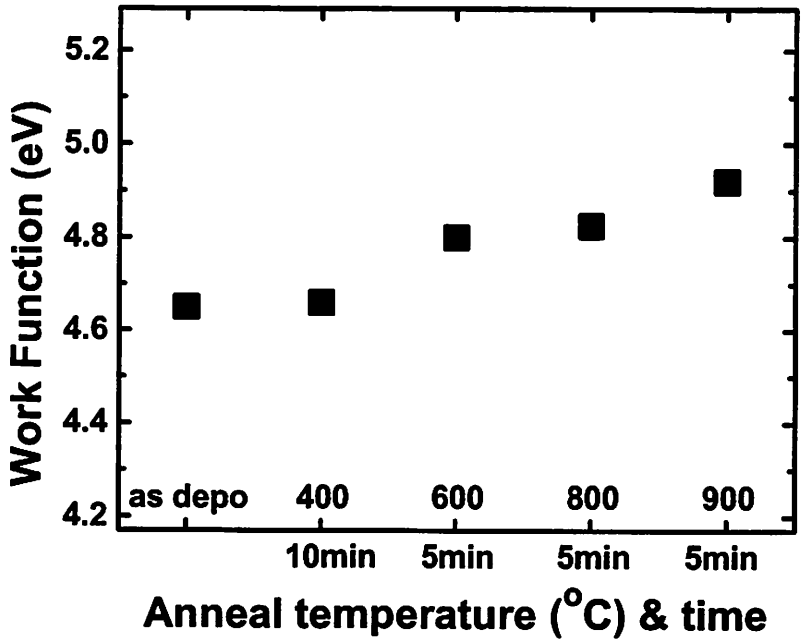


Fig. 5.19 The work function of Nb/Mo stack gate electrode changes from the value corresponding to (112) Nb to the value corresponding to (100) Mo.

TABLE 5.1. Work functions of selected elements. (Values are given for polycrystalline sample, and specific crystalline directions [5.11])

Element	Work function, (eV) / Crystalline direction	Element	Work function, (eV) / Crystalline direction
Ti	4.33	**Ni**	5.15
			5.22 (100)
Ta	4.25		5.04 (110)
	4.15 (100)		5.35 (111)
	4.80 (110)		
	4.00 (111)	**Ru**	4.71
Nb	4.3	**Mo**	4.6
	4.02 (001)		4.53 (100)
	4.87 (110)		4.95 (110)
	4.36 (111)		4.55 (111)
	4.63 (112)		4.36 (112)
	4.29 (113)		4.50 (114)
	3.95 (116)		4.55 (332)
	4.18 (310)		
Hf	3.9		
Zr	4.05		

5.7 Discussion

An advantageous dual work function metal gate fabrication process has been demonstrated. This process relies on metal interdiffusion to achieve the desirable work functions for NMOS and PMOS gate electrodes, and thus the gate dielectric is always protected from etchants. By eliminating the polysilicon depletion effect, metal gates reduce the total capacitance-equivalent thickness of the gate dielectrics, and as a result, increase the performance of MOS transistors.

The metal interdiffusion gate process can be implemented using a variety of metal systems. There are several factors which are going to dictate the choice of the metals to

be used. The transistor structure (bulk versus ultra-thin body) will determine the required work function of the metal gate electrodes. As demonstrated by our experiments, Ti and Ni have the work functions that correspond to the conduction and valence bands of silicon, and are therefore, good candidates for bulk CMOS applications. Nb and Mo have the work functions that are closer to the mid-band of silicon, and can therefore, be useful for ultra thin body (or double-gate) MOSFET applications.

The choice of the gate dielectric (SiO_2 versus high-κ) and the fabrication flow (gate-first versus gate-last) will also have an impact on the choice of the metals. It is well known that many of the low-work-function metals are not stable on SiO_2 during the high temperature source/drain activation step. Consequently, the transition to high-κ dielectrics, such as ZrO_2 or HfO_2, will increase the number of metals that can be used for NMOS gate electrodes.

Another promising pair of metals for bulk CMOS applications is Ta and Ru [5.12]. While Ta by itself (just like Ti) is not thermally stable on SiO_2, a Ta-Ru alloy is stable on SiO_2, as long as Ru concentration is larger than 40%. Importantly, the work function of the alloy remains low for Ru concentrations less than 54%. In reference 5.12, the NMOS Ta-Ru and the PMOS Ru gates were sputter-deposited on separate wafers. At that time, there was no proposal on how the two gates could be integrated into a CMOS process. The metal interdiffusion gate process proposed here can be used to facilitate the integration of Ta-Ru alloys into a manufacturable CMOS process.

Another advantage of miscible alloys, such as Ta-Ru, is that by continuously changing the composition of the alloy, the work function of the gate electrode can be continuously adjusted. This will make it possible to control the threshold voltage without the need to change the doping concentration in the channel of an MOS device.

Together with the use of high-κ dielectrics, MIG technology can provide the means for aggressive scaling of the gate dielectric capacitance-equivalent thickness for improvement of CMOS performance beyond the 50-nm technology node.

5.8 References

5.1 D. Kahng, M. M. Attala, "Silicon-silicon dioxide field induced surface devices," *in IRE-AIEE Solid-State Device Research Conference,* Pittsburgh, PA, 1960.

5.2 J. C. Sarace *et al.*, "Metal-nitride-oxide-silicon field effect transistors with self-aligned gate," *J. Solid-State Electron.,* Vol. 11, pp. 653-660, 1968.

5.3 F. M. Wanlass and C.T. Sah, "Nanowatt logic using field-effect metal-oxide semiconductor triodes," in *ISSCC Digest*, pp. 32-33, Feb. 1963.

5.4 C. H. Lee, H. F. Luan, W. P. Bai, S. J. Lee, T. S. Jeon, Y. Senzaki, D. Roberts, D. L. Kwong, "MOS characteristics of ultra thin rapid thermal CVD ZrO_2 and Zr silicate gate dielectrics," *International Electron Devices Meeting,* pp. 27-30, 2000.

5.5 H. F. Luan, B. Z. Wu, L. G. Kang, R. Vrtis, D. Roberts, D. L. Kwong, 'Ultra thin high quality Ta_2O_5 gate dielectric prepared by in-situ rapid thermal processing," *International Electron Devices Meeting*, pp. 609-612, 1998.

5.6 I. De, D. Johri, A. Srivastava, C. M. Osburn, "Impact of gate workfunction on device performance at the 50 nm technology node," *Solid-State Electronics*, Vol. 44, No.6, pp. 1077-1080, June 2000.

5.7 Xuejue Huang, Wen-Chin Lee, Charles Kuo, Digh Hisamoto, Leland Chang, Jakub Kedzierski, Erik Anderson, Hideki Takeuchi, Yang-Kyu Choi, Kazuya Asano, Vivek Subramanian, Tsu-Jae King, Jeffrey Bokor and Chenming Hu, "Sub-50 nm FinFET: PMOS," *International Electron Devices Meeting Technical Digest,* pp. 67-70, 1999.

5.8 Yang-Kyu Choi, K. Asano, N. Lindert, V. Subramanian, Tsu-Jae King, J. Bokor, Chenming Hu, "Ultra-thin body SOI MOSFET for deep-sub-tenth micron era," *International Electron Devices Meeting Technical Digest,* pp. 919-921, 1999.

5.9 L. Chang, S. Tang, T.-J. King, J. Bokor, and C. Hu, "Gate-Length Scaling and Threshold Voltage Control of Double-Gate MOSFETs," *International Electron Devices Meeting Technical Digest*, pp. 719-722, 2000.

5.10 K. Roy, R. Krishnammthy, "Design of low voltage CMOS circuits," *Tutorial Guide*, IEEE International Symposium on Circuits and Systems, Sydney, NSW, Australia, 6-9 May 2001.

5.11 Herbert B. Michellson, "The work function of the elements and its periodicity," *J. Appl. Phys.,* Vol. 48, No. 11, pp. 4729-4733, Nov. 1977.

5.12 Huicai Zhong, Greg Heuss, Veena Misra, "Electrical properties of RuO_2 gate electrodes for dual metal gate Si-CMOS," *IEEE Electron Device Letters*, Vol. 21, No.12, pp. 593-595, Dec. 2000.

5.13 Dae-Gyu Park, Tae-Ho Cha, Kwan-Yong Lim, Heung-Jae Cho, Tae-Kyun Kim, Se-Aug Jang, You-Seok Suh, Veena Misra, In-Seok Yeo, Jae-Sung Roh, Jin Won Park, Hee-Koo Yoon, "Robust ternary metal gate electrodes for dual gate CMOS devices," *International Electron Devices Meeting Technical Digest*, pp. 671-674, 2001.

5.14 Huicai Zhong, Shin-Nam Hong, You-Seok Suh, Heather Lazar, Greg Heuss and Veena Misra, "Properties of Ru-Ta alloys as gate electrodes for NMOS and PMOS silicon devices," *International Electron Devices Meeting Technical Digest*, pp. 467-470, 2001.

5.15 Y. H. Kim, C. H. Lee, T. S. Jeon, W. P. Bai, C. H. Choi, S. J. Lee, L. Xinjian, R. Clarks, D. Roberts, "High Quality CVD TaN Gate Electrode for Sub-100nm MOS Devices," *International Electron Devices Meeting Technical Digest*, pp. 667-670, 2001.

5.16 Y. C. Yeo, Q. Lu, P. Ranade, H. Takeuchi, K. J. Yang, I. Polishchuk, T.-J. King, C. Hu, S. C. Song, H. F. Luan, and D.-L. Kwong, "Dual-metal gate CMOS technology with ultra-thin silicon nitride gate dielectric", *IEEE Electron Device Lett.*, vol. 22, pp. 227-229, May 2001.

5.17 B. Yu, Y.-T. Tung, S. Tang, E. Hui, T.-J. King, C. Hu, "Ultra-thin-body silicon-on insulator MOSFET's for Terabit-scale integration," *Int. Semiconductor Device Research Symp.*, pp. 623-624, 1997.

5.18 E. H. Nicollian, J. R. Brews, MOS (Metal Oxide Semiconductor) Physics and Technology, p. 467, John Whley & Sons, New York, 1982.

5.19 M.-S. Liang, J. Y. Choi, P.-K. Ko, C. Hu, "Inversion-layer capacitance and mobility of very thin gate-oxide MOSFET's," *IEEE Trans. Electron Devices*, ED-33, pp. 409-413, March 1986.

5.20 Askill, Gibbs, "Tracer diffusion in β-titanium," *Phys. Status Solidi,* Vol. 11, p. 559, 1965.

5.21 G. B. Gibbs, D. Graham, D. H. Tomlin, Phil. Mag., **8**, p. 1269, 1963.

5.22 T. B. Massalski (editor-in-chief), Binary alloy phase diagrams, 2nd edition, ASM International, Materials Park, Ohio, p. 2875, 1990.

5.23 P. Ranade, H. Takeuchi, T. -J. King and C. Hu, "Work Function Engineering of Molybdenum Gate Electrodes by Nitrogen Implantation," Electrochemical and Solid State Letters, Vol. 4, No. 11, pp. G85-G87, Nov. 2001.

Chapter 6 Conclusion

6.1 Summary

We have considered several important issues related to fabrication, modeling, and reliability of gate stacks for MOSFET devices in this dissertation.

A detailed reliability study of 100 nm MOS transistors with 14Å-equivalent-oxide-thickness JVD silicon nitride gate dielectric was presented in Chapter 2. The study shows that a hard dielectric breakdown due to Folwer-Nordheim stress is no longer the major reliability concern for this ultra-thin gate dielectric. Instead, a gradual degradation of the device parameters such as carrier mobility, threshold voltage, and gate leakage current can become the limiting factor that determines the useful lifetime of a semiconductor product. A set of empirical equations that describe this degradation as a function of time and stress voltage was proposed. This set of equations can be used to predict the amount of device degradation under various operating conditions. The hot-carrier reliability study revealed that the degradation mechanism and expected lifetime are same for PMOSFETs with silicon nitride and silicon oxide gate dielectrics. Thus from the reliability perspective JVD silicon nitride is a viable successor to silicon oxide.

The electron tunneling through multi-layer gate dielectrics was modeled in Chapter 3. Based on this developed model the tunneling current through both single- and multi-layer dielectrics can be closely approximated as an exponential function of equivalent oxide thickness. Consequently, leakage current through any dielectric stack can be determined based on a single number: the tunneling attenuation coefficient.

A new interface scattering model based on carrier wavefunction penetration into gate dielectric was proposed in Chapter 4. This model is an alternative to the surface roughness model and is in a better agreement with the experimental data. When the interface scattering is combined with Coulombic and phonon scattering, the mobility data for a variety of gate dielectrics and the full range of gate bias conditions can be matched.

An interdiffusion-based method for fabrication of dual work function metal gate electrodes for CMOS devices was proposed in Chapter 5. The feasibility of metal interdiffusion gate (MIG) process was demonstrated by making CMOS transistor with Ti and Ni gates. The work functions of Ti and Ni correspond to the conduction and valence bands of silicon, making this metal system suitable for bulk CMOS applications. It has also been shown that other metal systems (such as Nb-Mo) can be used to implement the MIG process. Therefore a wide range of gate electrode work functions can be potentially achieved to fit the requirements for various CMOS device structures.

6.2 Contributions

The research findings presented in this dissertation has made several significant contributions in the area of gate stack engineering for sub-0.1μm CMOS devices:

The reliability study for Si_3N_4 has proven that this material is a viable gate dielectric for at least several generations of CMOS technology. Furthermore, the methods and models used to evaluate the reliability of silicon nitride can be used as a stepping-stone towards an understanding of the reliability of true high-κ dielectrics.

The proposed analytical tunneling model for multi-layer gate dielectrics is an important tool for understanding the scaling limits of various gate dielectric stacks. According to the model there is a family of universal current-versus-thickness lines that describes the scaling behavior for any dielectric stack. Therefore, the scaling behavior for multi-layer dielectrics is essentially the same as for single-layer dielectrics.

The new interface scattering model for channel carriers allows the extension of the universal mobility model from SiO_2 to high-k gate dielectrics. The model also helps us understand how a choice of the gate dielectric for future MOS transistors would affect surface channel mobility.

Finally, the proposed metal interdiffusion gate process helps to resolve the critical challenge of integrating different-work-function metals on a single CMOS wafer while protecting the integrity of the gate dielectric.

These contributions will facilitate a successful integration of such new materials such as high-κ dielectrics and metal gates into a manufacturable CMOS process, and therefore extend the scaling limits of Si-based semiconductor technology.

6.3 Suggestions

6.3.1 Reliability of high-κ dielectrics

It might be necessary to have a high-κ gate dielectrics ready for production as soon as year 2005 in order to meet the leakage current specifications for some low power IC applications. Clearly, no material can be used in production until its reliability is established. Therefore there is a pressing need to study the reliability of promising high-κ dielectrics such as HfO_2. This work is still nascent. While it appears that HfO_2 is sufficiently immune to hot-carrier degradation [6.1], the detailed studies of HfO_2 reliability under Folwer-Nordheim stress have only been performed on capacitors [6.2, 6.3]. In order to investigate the impact of Folwer-Nordheim stress on gradual deterioration of MOS devices, reliability studies should be performed on MOS transistors (similar to work presented in Chapter 2, section 3).

6.3.2 Interfacial layers: Impact on tunneling and mobility

In this dissertation we addressed the issue of the interfacial dielectric layers twice: in our discussion of tunneling, and in our discussion of channel mobility. In both cases, we modeled these layers as rectangular (or trapezoidal, when a voltage is applied) barriers. This means that we assumed an absolutely abrupt edge between the interfacial layer and Si substrate, as well as constant values for barrier height and effective electron mass throughout the interfacial layer. These assumptions would be reasonable for dielectric layer with thickness much larger than the inter-atomic spacing. However if a

layer is only several angstroms thick, it becomes necessary to account for the fact the dielectric band-gap properties vary gradually near the interfaces [6.4].

The understanding of the electrical properties is especially important to precisely determine the amount of wavefunction penetration and its impact on carrier mobility. The sssumption that the electrical properties of a SiO_2 interfacial layer are the same as those of bulk SiO_2 would lead to an erroneous conclusion that a 2Å SiO_2 layer would be sufficient to ensure good mobility. In realty, a thicker layer (perhaps 4 to 7 Å) is required. Since theoretical band structure calculations [6.4] are often imprecise, experimental methods are preferred in the analysis of the properties of interfacial layers. High-resolution electron energy-loss spectroscopy (EELS) is arguably the best technique to determine the chemical composition and electronic structure of the interfacial layers.

Same comments apply to modeling of gate tunneling. While it still should be possible to apply the concept of tunneling attenuation coefficient α to model the gate leakage current, the question of finding the value of α for stacks with gradually varying interfacial properties still has to be addressed.

6.3.3 Epitaxial gate dielectrics

The beneficial effect that epitaxial gate dielectrics would have on MOSFET channel carrier mobility was described in Chapter 4. A number of high-κ dielectrics can be grown epitaxially on (100) Si: $SrTiO_3$ [6.5], CeO_2 [6.6], Y_3O_2 [6.7], and yittria-stabilized-zirconia. It still remains to be clearly demonstrated that MOSFETs with these gate dielectrics can be successfully fabricated. It is quite a challenge to avoid the formation of a thin amorphous layer between the crystalline dielectric and Si substrate. A

presence of such an amorphous layer would clearly negate the mobility benefits that epitaxial dielectrics are expected to provide.

In addition to improved carrier mobility the epitaxial gate dielectrics have other potential advantages for MOSFET integration. For example, it is possible to grow a single-crystalline silicon gate on top of these dielectrics. Dopant activation in a single-crystalline gate can be higher than in poly-crystalline one. Hence the polysilicon depletion effect can be suppressed. Furthermore, it would become possible to make structures with multiple alternating single-crystalline layers of semiconductors and dielectrics (Fig 6.1). These structures could then be potentially used to make "multi-channel" FETs. The structure shown in Figure 6.1 can also facilitate the 3-D integration of integrated circuits.

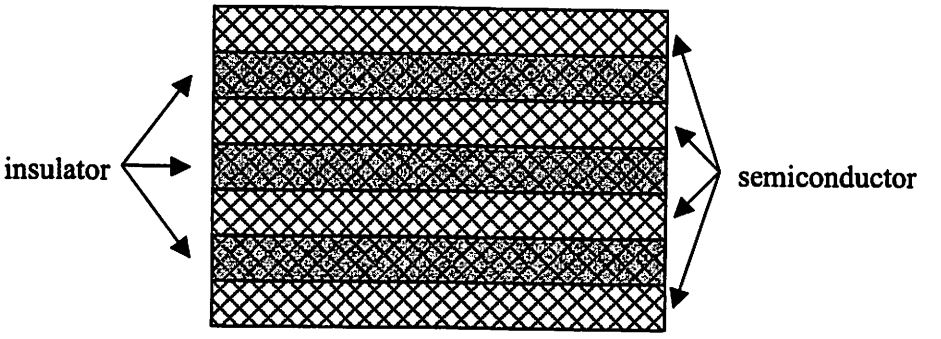


Fig. 6.1. A semiconductor-insulator epitaxial structure can help to make multi-channel transistors and achieve 3-D integration.

6.3.4 Mobility modeling for ultra-thin body transistors

In our discussion of carrier channel mobility (Chapter 4), we considered how two out of three scattering components (Coulombic and interfacial) can change when a different gate dielectric is introduced. Phonon scattering however, should remain

essentially unchanged as long as the channel material remains the same. Clearly if the transistor channel is made of SiGe instead of Si, or strained Si, then the changes in the band structure will affect the phonon scattering rates.

Interestingly enough, even in the case when transistor channel is made of relaxed Si, the phonon scattering rates might change if the thickness of this silicon layer is reduced sufficiently. Several factors can cause this change. First of all, depending on the Si body thickness and gate bias, the inversion layer carriers are not necessarily going to be confined by the vertical electric field. They may also be confined by the thickness of the silicon body itself. In other words, the band structure (or more specifically the sub-band splitting) would be modified due to the finite thickness of the silicon body. Consequently the phonon scattering rates will change. Furthermore, the phonon dispersion also changes when the silicon body is less then 10nm thick [6.8]. This is another reason the phonon scattering rates might change.

These effects related to the 2-dimensional nature of the transistor body will certainly have to be examined in order to understand the channel carrier mobility in ultra-thin body devices.

6.3.5 Gate work function engineering

The major challenge for metal gate CMOS is finding a stable metal with low work function. A number of potential ways to overcome this limitation can be explored in the future.

First, a low work function metal can be alloyed with a high work function metal to improve the NMOS gate electrode thermal stability while retaining a relatively low

work function. It has been shown that there is at least one metal system (Ru-Ta [6.9]) that has these desired properties.

Second, not only pure metals or metal alloys, but also metal silicides can be considered as potential candidates for gate electrodes. Silicides have mobile carrier concentrations high enough to successfully suppress the gate depletion effect. The advantage of such a process is that it is very similar to the salicide process currently used. In a salicide process metal silicide is formed in the top portion of the gate electrode. In order to implement a silicide gate process the front of the silicidation reaction should be extended all the way to the gate dielectric interface. Unfortunately, it has been difficult to find metal silicides with work functions far enough from the Si mid-gap to be suitable for bulk CMOS device. Therefore the use of silicide gates can be limited to ultra-thin body applications.

Finally, we shall point out that it is especially hard to fabricate PMOSFET polysilicon gates with high doping concentration (due to the fact that boron diffuses easily through thin gate dielectrics). At the same time, a very high doping concentration can be achieved in NMOSFET polysilicon gates (via in-situ doping). Therefore combining N-doped polysilicon gate with a high-work-function metal (or silicide) gate can be an attractive way to suppress the polysilicon depletion effect without the need for reactive low-work-function metals.

6.4 References

6.1 Qiang Lu, Hideki Takeuchi, Ronald Lin, Tsu-Jae King, Chenming Hu, Katsunori Onishi, Rino Choi, Chang-Seok Kang, Jack C. Lee, "Hot Carrier Reliability of n-MOSFET with Ultra-thin HfO_2 Gate Dielectric and Poly-Si Gate," *in Proceedings of Int. Reliability Phys. Symp.*, pp. 429-431, Dallax, TX, Apr. 2002.

6.2 S. J. Lee, C. H. Lee, C. H. Choi, D. L. Kwong, "Time-dependent Dielectric Breakdown in poly-Si CVD HfO_2 Gate Stack," *in Proceedings of Int. Reliability Phys. Symp.*, pp. 409-414, Dallax, TX, Apr. 2002.

6.3 Katsunori Onishi, Chang Seok Kang, Rino Choi, Hag-Ju Cho, Sundar Gopalan, Renee Nieh, Siddharth Krishnan, Jack C.Lee, "Charging Effects on Reliability of HfO_2 Devices with Polysilicon Gate Electrode," *in Proceedings of Int. Reliability Phys. Symp.*, pp. 419-420, Dallax, TX, Apr. 2002.

6.4 Takahiro Yamasaki, Chioko Kaneta, "Geometric and electronic structures of SiO_2/Si(001) interfaces," *Phys. Rev. B,* Vol. 63 pp. 115314-1 – 115314-5, 2001.

6.5 R. A. McKee, F. J. Walker, M. F. Chisholm, " Crystalline oxides on silicon: the first five monolayers," *Physical Review Letters*, Vol. 81, No.14, pp. 3014-3017, Oct. 1998.

6.6 Tomoyasu Inoue, Tetsu Ohsuna, Yasuhiro Obara, Yasuhiro Yamamoto, Masataka Satoh, Yoshinobu Sakurai, "Intermediate Amorphous Layer Formation Mechanism at the Interface of Epitaxial CeO_2 Layers and Si Substrates," *Jpn. J. Appl. Phys.*, Vol. 32, No. 4, pp. 1765-1767, 15 Apr. 1993.

6.7 S. C. Choi, M. H. Cho, S. W. Whangbo, C. N. Whang, S. B. Kang, S. I. Lee, M. Y. Lee, "Epitaxial growth of Y2O3 films on Si(100) without an interfacial oxide layer," *Appl. Phys. Lett.*, Vol. 71, No.7, pp. 903-905, 18 Aug. 1997.

6.8 A. A. Balandin, "Nanoscale thermal management," *IEEE Potentials*, Vol. 21, No.1, pp. 11-15, Feb.-Mar. 2002.

6.9 Huicai Zhong, Shin-Nam Hong, You-Seok Suh, Heather Lazar, Greg Heuss and Veena Misra, "Properties of Ru-Ta alloys as gate electrodes for NMOS and PMOS silicon devices," *IEDM*, pp. 467-470, 2001.