Title

Reliability and Performance Enhancement in Ultra-Scaled Advanced CMOS and Memory Devices

ISBN

9798189271342

Author

Lee, Dasom

Publication Date

2026-08-01

Peer reviewed|Thesis/dissertation

Reliability and Performance Enhancement in Ultra-Scaled Advanced CMOS and Memory Devices

by

Dasom Lee

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 Tsu-Jae King Liu, Chair
Professor Sayeef Salahuddin
Professor Oscar Dubon

Summer 2026

Reliability and Performance Enhancement in Ultra-Scaled Advanced CMOS and Memory Devices

Dasom Lee

Abstract

Reliability and Performance Enhancement in Ultra-Scaled Advanced CMOS and Memory Devices

by

Dasom Lee

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

To CJ and Roy,

for their love

Contents

List of Figures

List of Tables

Acknowledgments

As the saying goes, "It takes a village to raise a child," and throughout my Ph.D. journey, the entire semiconductor community has been that village for me.

First and foremost, I would like to express my gratitude to my advisor, Prof. Tsu-Jae King Liu. Through her example, I learned how far one's impact can reach in both academia and industry when knowledge is combined with integrity and wisdom. She taught me not only the fundamentals of research, but also the attitude needed to navigate both work and life. She provided an exceptional research environment where I was able to explore the academic world freely while always feeling supported and guided. I am especially grateful for her patience and unwavering encouragement through the many failures and challenges in my research projects. From her, I learned that the true nature of research lies not in avoiding failure, but in having the perseverance to continue despite it. It will always be one of the greatest honors of my life to have been her student.

I would also like to extend my gratitude to my dissertation committee. Professor Sayeef Salahuddin not only served on my dissertation committee, but also entrusted me with the role of head TA for EECS 16B. Through his mentorship, I gained invaluable knowledge and learned the true meaning of leadership. In my research, he always provided practical advice and effective solutions whenever challenges arose. I am grateful for his guidance and support throughout my Ph.D. Professor Oscar Dubon also served on my dissertation committee. I appreciate his interest in and support for my research. His perspective as a materials science professor helped me approach my research with a broader and deeper understanding. Lastly, I would like to thank Dr. Tahir Ghani, who served on my qualification committee as a visiting professor and Intel Senior Fellow, for mentoring me throughout the last two years. It was a tremendous privilege to receive his personalized guidance on a weekly basis and learn how to approach research thoughtfully. His passion for research and depth of insight have been a constant source of inspiration throughout my Ph.D.

My research was supported by several companies, and I appreciate for their generous support. First of all, I would like to thank SK hynix for making it possible for me to pursue a Ph.D. I never imagined I would be given a second opportunity to return to academia. My years at SK hynix before beginning my doctoral studies gave me a clear understanding of what I needed to learn, allowing me to approach both my coursework and research with purpose and direction. The RTN project was supported by Samsung Electronics through the provision of FinFET devices. I am grateful to Dr. Yuri Masuoka and Donghoon Kim for their continuous collaboration and technical support, which were instrumental in advancing this work. I would also like to thank Professor Hiu Yung Wong at San Jose State University for his valuable insights and guidance on the RTN simulation. The OI (oxygen-insertion layer) project was sponsored by Atomera. In particular, I am sincerely grateful to Hideki Takeuchi, Vice President of Engineering at Atomera, who mentored me throughout the project. He was always approachable whenever I needed guidance. Working with Hideki reminded me

how enjoyable research can be, especially when collaborating with someone who genuinely loves technology and scientific discovery.

My fabrication experience at Berkeley was challenging at first, and it would not have been possible without the help of the Nanolab staff. I am thankful for Michael Helmbrecht, Dave Taosaka, Sam Tsitrin, Ryan Rivers, Tariq Haniff, Erin Hancock, Sanchay Gadia, Gabriel Flores, Allison Dove, Anthony Bielecki, Joanna Bettinger, David Lo, Jesse Merz, Greg Mullins, and Daniel Martinez for their assistance and guidance throughout my work. I would also like to thank Ashwini Gopal, Khue Ngo, Ivy Nguyen, and Emmuelle Marie Ejanda from Nanoshift for their practical help and advice whenever I encountered challenges. I am especially grateful to Ashwini for her encouragement and for reminding me that learning is always part of the process. Finally, I would like to thank Kyungmok Kwon from nEye Systems, Inc. for his generous training and continued support.

Studying at Berkeley gave me many unforgettable moments, thanks to the wonderful colleagues who became an important part of my life. From the time I joined the group, Dr. Lars Tatum and Dr. Urmita Sikder were not only colleagues but also friends who became like family to me. They helped me adapt to an unfamiliar environment and made my transition to Berkeley much smoother. Dr. Lars Tatum helped me establish my research environment, while Dr. Urmita Sikder always offered practical and honest advice whenever I faced challenges in the lab. They also encouraged me to step outside of school and experience more of the world beyond academia. Their warmth and kindness profoundly shaped me and made me a better person. Dr. Tsegereda Kedir Esatu and Dr. Minwu Kim also warmly welcomed me when I first joined the group and helped me feel at home in Berkeley. In my second year, Collin Patrick Finnan joined the group. His passion and creativity continually inspired me in research, and his kind words always comforted me during difficult times. I feel fortunate to have shared my years at Berkeley with him. Dr. Choong-Ho Rhee joined the group as a visiting scholar, and his warm and thoughtful personality always brought positive energy to everyone around him. Dr. Donghyun Ryu, who recently joined the group as a postdoctoral researcher, became a wonderful discussion partner. He adapted quickly to the group and was always eager to contribute, making him a valued member of our team.

Beyond my group members, it would not have been possible for me to complete my Ph.D. without my friends Jen-Hao Chen and Chirag Garg. We took most of our classes together and went through the preliminary exam and qualification process at around the same time. I was grateful to have friends who understood my situation and were always willing to help throughout every stage of my Ph.D. Their ability to learn quickly, their strong work ethic, and their dedication continually impressed me, and I have no doubt that they will become engineers who shape the future of technology.

I extend my gratitude to the members of Prof. Sayeef Salahuddin's group. They were wonderful companions while we shared an office in Sutardja Dai Hall and worked on similar research topics. I sincerely thank Dr. Nirmaan Shanker, who trained me on the measurement tools in the Device Characterization Lab. He believed it was his responsibility to guide new

students and take the best care of the lab equipment. His thoughtful training gave me confidence and helped me overcome my fear of using unfamiliar tools and processes. My sincere thanks also go to Koushik Das, whose expertise in film deposition guided me in optimizing my high-k/metal gate stacks. Without his guidance, it would have taken much longer to stabilize and refine the process. He generously shared his knowledge and experience without hesitation. I am also thankful to Dr. Jason Hsu, Professor Suraj Cheema, Dr. Li-Chen Wang, Dr. Ratik Brahma, Steve Volkman, Dr. Adi Jung, Dr. Jongho Park, Namho Jeon, Dr. Seungho Yang, Junhao Chen, Chia-Chun Lee, Richard McManus, Wei-Yang Weng, Kim Pham, Krishna Bhattaram, Aniket Sadashiva, Ramit Dutta, Alex Schmidt, Professor Asir Khan, Dr. Myeongseop Song, Dr. Junghyeon Hwang, Jooseong Yun, and Yonghyeok Son, who trained me on many other tools and were wonderful colleagues throughout my time at Berkeley. Through them, I experienced the strong sense of community and collaboration that makes the semiconductor device field so special.

Outside of school, I spent most of my time in UC Village. As the mother of a young child, UC Village became a place I could always rely on. Jieun and Kyoungsuk often shared homemade meals with me, especially since cooking was never my strength. What they gave me was more than food—it was a warm sense of belonging and community. Jeongmin shared valuable advice about raising children, and our conversations always brought me comfort and renewed my energy. After my husband returned to Korea, Daun and Seungpyo looked after Roy and me with great kindness. Daun always created joyful activities that brought everyone together, and I hope she continues to enjoy a happy life in the village after I leave. Lastly, I am deeply thankful to Surin and Sanghoon for their friendship. Spending time with young children is not always easy, yet they were always patient, thoughtful, and kind to Roy and me. I hope I will have the opportunity to return their kindness by caring for their child someday.

During my time at Berkeley, visiting Korea gave me a refreshing break and a chance to recharge. Knowing that I was studying abroad, my friends always made time for me despite my limited schedule. They took time off from work, came to wherever I was staying, and prepared delicious meals and thoughtful gifts that reminded me of home and the things I had missed. I am fortunate to have friends like Jooran, Gyuri, Harim, Min, Heejin, Cholong, Yeoulmae, and Bokyung who are are thoughtful and warmhearted. I am especially grateful to Boryeon, who visited me in Berkeley. Spending those two weeks together remains one of my most treasured memories.

Finally, I would like to express my deepest gratitude to my family. Earning a Ph.D. at Berkeley required sacrifices from every member of my family, and this journey would not have been possible without their unwavering love and support. First of all, my son Roy came to the United States when he was only 18 months old and spent four years of his childhood growing up here. He experienced long periods of separation from his dad and the challenges of adapting to a new culture and environment. I am deeply grateful for how well he adapted to these changes and for the loving relationships he built with the many caregivers who

looked after him. My husband, CJ, took two years of paternity leave to support our family throughout my Ph.D. Without his unconditional love, patience, and sacrifice, I could never have begun or completed this Ph.D. Lastly, my mom and dad came to the United States several times to care for Roy. They set aside their own lives in Korea and stayed here simply to support Roy and me. Their love and dedication gave me the strength to keep moving forward. My family has always been my greatest source of strength, motivation, and inspiration. This Ph.D. is as much theirs as it is mine, and I hope this achievement reflects even a small part of the sacrifices they made for me.

Chapter 1

Introduction

1.1 Emerging AI Computing

Artificial intelligence (AI), particularly large language model (LLM)-based applications, has dramatically increased the demand for computing [1]. The computational resources required for AI applications continue to grow rapidly as model sizes and data volumes increase. Consequently, future computing systems require evolution in computing architectures to improve data processing capability [2].

Figure 1.1 (a) illustrates the conventional von Neumann computer architecture in which the central processing unit (CPU) and memory unit are implemented on separate semiconductor substrates ("chips"), so they are physically separated, and data is transferred between them via a bus comprising metallic wires interconnecting the chips. For AI applications, data transfer becomes a major performance bottleneck because large amounts of data must be continuously exchanged between the processor and memory during training and inference operations. Therefore, reducing communication overhead between processing and memory units has become one of the key challenges for next-generation computing systems.

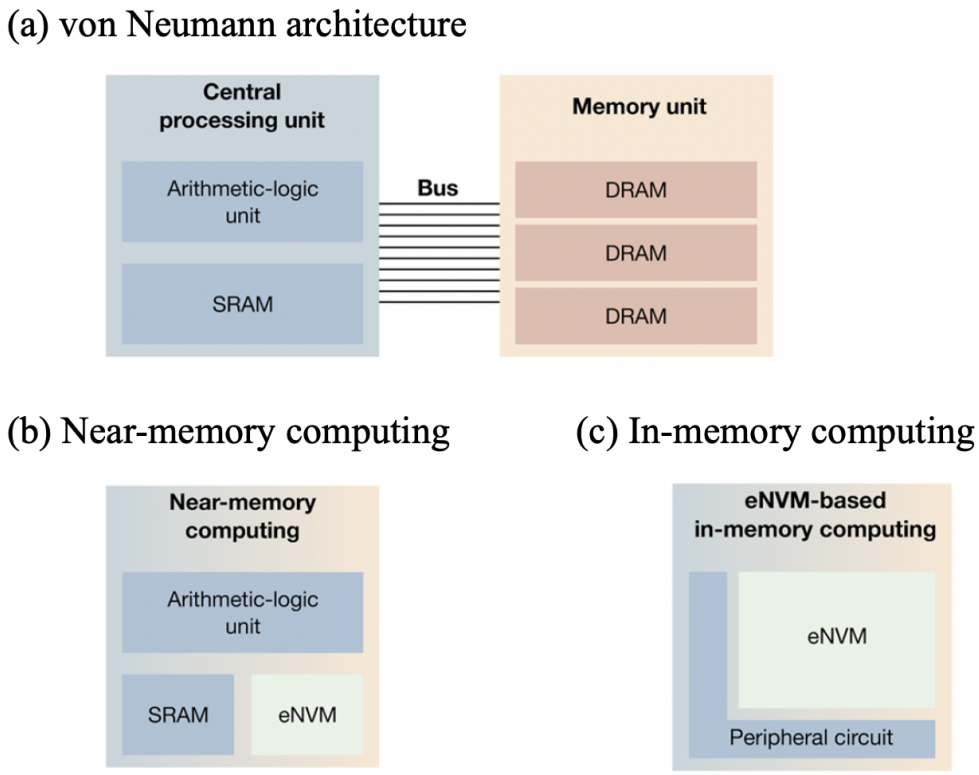


Figure 1.1: Computing architectures adapted from [3]. (a) Conventional von Neumann architecture, (b) near-memory computing architecture, and (c) in-memory computing architecture based on emerging nonvolatile memory (eNVM).

To alleviate the data-transfer bottleneck, near-memory computing (Figure 1.1 (b)), and in-memory computing (Figure 1.1 (c)) architectures have been actively investigated [3]. In near-memory computing systems, memory and computational units are placed physically closer together to minimize communication overhead [4]. In contrast, in-memory computing performs certain computational operations directly within the memory array itself [5]. Such architectures are particularly suitable for neural-network-based AI applications, where repeated matrix operations and large-scale data movement dominate the computational workload.

In addition to these architectural approaches, advanced memory technologies have also been developed to increase data transfer throughput. High-bandwidth memory (HBM) [6] improves communication between processors and DRAM through three-dimensional (3D) stacking, while high-bandwidth flash (HBF) memory applies concepts similarly as 3D NAND flash memory to enable higher-bandwidth data storage. Together, these technologies enhance

both volatile and non-volatile memory performance, to meet increasing data-processing demand.

Although these architectural innovations improve the efficiency of data transfer, the overall performance of AI systems ultimately depends on the performance and integration density of the individual transistors and memory devices, i.e. the basic integrated circuit (IC) building blocks. Therefore, continued advances in semiconductor device technologies remain essential for meeting the increasing computational demands of future AI systems. However, as illustrated in Figure 1.2, transistor density scaling has slowed considerably beyond the 5 nm technology node because of fabrication complexity and the difficulty of maintaining desirable electrical characteristics in short-channel devices [7]. As a result, sustaining future computing performance requires innovations in semiconductor materials, device structures, and fabrication technologies that enable higher integration density while improving energy efficiency.

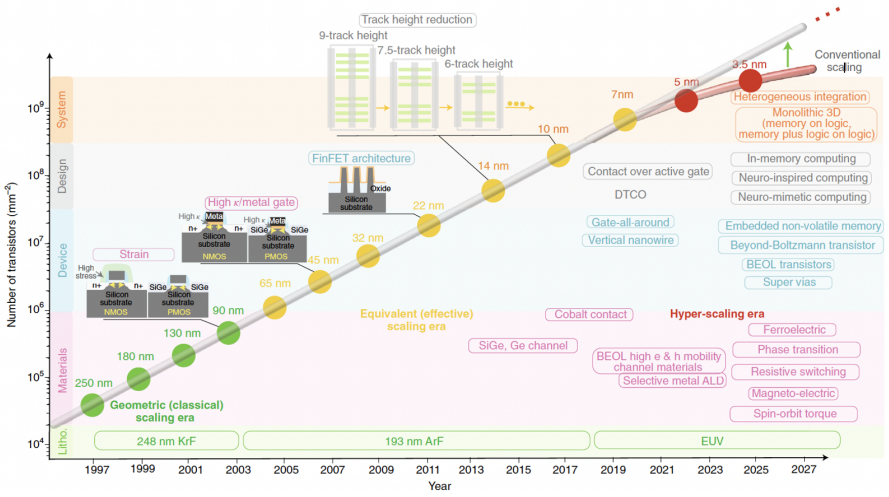


Figure 1.2: CMOS technology scaling trend adapted from [7].

1.2 Scaling Challenges for Advanced CMOS Devices

One of the major challenges associated with further miniaturization of complementary metal-oxide-semiconductor (CMOS) transistors, the predominant technology used for digital computing, is the short-channel effect [8, 9], in which the capacitive coupling between the drain and the channel region near the source increases as the gate length decreases. This results in degraded gate control which worsens the tradeoff between high on-state current and low

off-state current.

Gate controllability is commonly gauged by the subthreshold swing (SS), which represents the change in gate voltage required to change the drain current by one order of magnitude in the subthreshold region of operation [10]. A steeper switching characteristic, corresponding to smaller SS, enables lower threshold voltage while maintaining low off-state leakage current (I_{OFF}); this in turn allows for lower IC operating voltage (V_{DD}) to achieve the specified on-state current (I_{ON}), for reduced energy consumption.

The subthreshold swing is determined by several parameters, as shown in Eq. (1.1). Here, C_{ox} is the gate-oxide capacitance, C_{dep} is the depletion capacitance, k_B is the Boltzmann constant, T is the operating temperature, and q is the elementary charge.

$$SS = \left(1 + \frac{C_{\mathrm{dep}}}{C_{\mathrm{ox}}}\right) \frac{2.3\, k_{\mathrm{B}} T}{q} \tag{1.1}$$

One approach to improve subthreshold swing is to increase the gate oxide capacitance (C_{ox}) by reducing the equivalent gate-oxide thickness (EOT). Early CMOS technologies employed heavily doped polycrystalline silicon (poly-Si) as the gate electrode material together with silicon dioxide (SiO_2) as the gate oxide material [11]. However, a depletion layer formed within the poly-Si gate in the on-state effectively increases the EOT. Also, aggressively reducing the physical thickness of SiO_2 results in excessive gate leakage current and degraded oxide reliability. To overcome these limitations, dielectric materials with higher relative permittivity (high-k) together with metallic gate materials (HKMG) were introduced to increase C_{ox}. Today, hafnium dioxide (HfO_2) is the predominant (high-k) gate-oxide material used in leading-edge CMOS technologies [12]. The higher dielectric constant of HfO_2 enables a physically thicker gate-insulating layer for a given low value of EOT, thereby suppressing gate leakage current without sacrificing gate controllability. Metallic gate electrode materials do not have the gate depletion effect. However, they present a challenge for gate work function tuning to set the transistor threshold voltage. Metal gate material engineering and unintended interdiffusion between layers within the metal gate film remains a HKMG process integration challenge for advanced (fully depleted) transistor designs. Typically, to achieve a low effective gate work function, n-channel (NMOS) transistors incorporate an ultrathin aluminum (Al) layer sandwiched between titanium-nitride (TiN) layers to suppress Al diffusion [13].

To further improve gate controllability, thin-body (nanometer-scale) transistor structures such as FinFETs [14] have been adopted for sub-22 nm CMOS technology generations ("nodes"). By making the semiconductor ultra-thin, sub-surface off-state leakage current paths are eliminated and the depletion capacitance (C_{dep}) is reduced. Furthermore, by having the HKMG stack straddle the semiconductor channel region (as in the FinFET structure), gate-to-channel capacitive coupling is enhanced. The most advanced transistor structures used for high-volume chip production today are gate-all-around FETs (GAAFETs) [15], in

which the HKMG stack wraps around four sides of the thin-body semiconductor channel region. To provide for higher on-state drive current per unit area, multiple channel regions are stacked over each other, with two HKMG stacks back-to-back in-between them. For ease of fabrication (i.e., for higher manufacturing yield), the spacing between adjacent semiconductor "nanosheets" is limited. Therefore, alternative approaches to tuning the effective gate work function, which do not require a thick gate film, are of interest for future generations of CMOS technology.

The SS is limited to be no smaller than 60 mV/dec at room temperature, due to the Boltzmann distribution (exponentially decreasing number with increasing energy) of conduction electrons within the transistor source region [16]. To achieve even steeper switching characteristics (SS ¡ 60 mV/dec), the operating temperature can be reduced [17]. Cryogenic CMOS IC operation has attracted increasing interest for future energy-efficient computing systems. However, electrical behavior of advanced transistor structures under cryogenic operation are not yet fully understood. In particular, as transistor miniaturization (gate-length scaling) continues, the influence of individual defects becomes more significant, making reliability analysis essential for advanced CMOS technologies operating at cryogenic temperatures.

In summary, improvements are needed for tuning the effective gate work function and for ensuring reliable transistor operation, to facilitate continued scaling of advanced CMOS devices.

1.3 Scaling Challenges for Advanced Memory Devices

Semiconductor memory technologies also face significant challenges for continued scaling to increase data storage density. For dynamic memory (DRAM) bit-cells, maintaining sufficient storage capacitance becomes increasingly difficult as bit-cell dimensions shrink, resulting in reduced signal margin and degraded data retention. To address this limitation, various 3D-integrated DRAM architectures have been proposed, in which storage capacitors are stacked to increase the effective capacitor volume while maintaining a small footprint per bit-cell [18, 19]. As an alternative scaling strategy, capacitorless DRAM bit-cell designs have also attracted considerable attention [20, 21]. Capacitorless DRAM stores information within the access transistor body region and thereby eliminates the need for a separate capacitor, simplifying the bit-cell structure. However, technical challenges including reliable data storage and readout still need to be addressed before capacitorless DRAM can be widely adopted.

High-density non-volatile memory (NVM) with cells arranged in a NAND array architecture and which can be erased, written, and read in blocks ("flash" operations) transitioned from planar 2D arrays to vertically integrated 3D arrays more than a decade ago to enable storage density to increase beyond the limits of lateral bit-cell scaling [22]. For every vertical string of series-connected cells in a 3D NAND Flash memory array, the cells are formed using the same conformally and sequentially deposited blocking-oxide, charge-trap nitride,

tunneling-oxide (ONO) and semiconductor-channel layers within holes formed through a stack of planar individual gate electrodes and insulating oxide layers; thus each bit-cell gate stack (comprising the gate electrode and ONO layers) wraps around the semiconductor layer for independent local control of the channel electric potential.

Increasing the number of vertically stacked memory cells improves storage capacity for a given 3D NAND array footprint. However, higher aspect ratio of the holes formed through the gate electrode stack presents fabrication challenges and also can lead to degraded performance. As the stack height increases, nonuniform layer thickness along the vertical direction can cause variations in bit-cell program and erase characteristics, resulting in wider threshold voltage distribution, making it more difficult to store multiple bits within each cell.

In addition to device-to-device variability, the unique string architecture of 3D NAND flash memory introduces challenges for erase operation: hundreds of cells share a thin, vertical semiconductor channel layer with only one heavily doped drain junction located at the top of each string; to erase all the cells within a string, holes are generated in the channel layer near the drain junction by gate-induced drain leakage (GIDL) and they must propagate down the entire semiconductor layer to raise the local channel potential for each cell [23]. As the number of cells in the string increases, this hole transport becomes less efficient, resulting in slower erase operation. Unlike variability issues, which can often be mitigated through process optimization, this erase limitation is inherent to the 3D NAND architecture. Therefore, improving erase performance requires innovations at the device level.

In summary, advances in 3D NAND flash memory device design are needed to facilitate continued scaling of storage density while meeting performance specifications.

1.4 Outline of the Dissertation

This dissertation investigates reliability and performance enhancement techniques for advanced CMOS and memory devices to enable continued improvement in computing performance and energy efficiency. The work covers defect-induced transistor reliability phenomena, effective gate work function engineering, and advanced 3D NAND Flash device design.

Chapter 2 investigates random telegraph noise (RTN) in advanced MOSFETs operating at cryogenic temperatures. Experimental measurements using Samsung's 14nm FinFETs are combined with technology computer-aided design (TCAD) simulations to analyze the dependence of RTN characteristics on temperature and operating voltage conditions. In addition, RTN behavior in advanced transistor structures, including FinFETs and gate-all-around FETs (GAAFETs), are compared to project the impact of device scaling on defect-induced noise for cryogenic operation.

Chapter 3 investigates the effects of ultrathin aluminum incorporation within HKMG stacks with HfO_2 gate oxide. The impacts of Al on the effective gate work function and gate-oxide reliability are investigated through electrical characterization and material analyses.

In addition, the impact of post-HKMG forming gas anneal (FGA) is discussed.

Chapter 4 discusses oxygen insertion (OI) technology as an alternative approach for effective gate work function engineering in advanced NMOS devices. Unlike conventional work function engineering methods based on metal-layer intermixing, the OI approach suppresses intermixing while simultaneously enhancing electron mobility. This chapter introduces the OI technology concept and investigates its effectiveness for gate work function reduction through electrical characterization and chemical analysis.

Chapter 5 focuses on erase performance enhancement of 3D NAND flash memory. A SiGe/Si heterojunction drain design is proposed to significantly enhance gate-induced drain leakage (GIDL) for faster erase operation without degrading inhibit mode operation. TCAD simulations are performed to evaluate the effectiveness of the proposed design by analyzing band-to-band tunneling, erase efficiency, and inhibit mode operation in advanced 3D NAND structures.

Finally, Chapter 6 summarizes the major contributions of this dissertation and discusses future research directions for advanced CMOS and memory technologies.

Chapter 2

Study of Defect-Induced Noise in Advanced MOSFETs for Cryogenic Operation

2.1 Introduction

Metal-oxide-semiconductor (MOS) field-effect transistors (FETs) exhibit significantly enhanced on-state "drive" current at cryogenic temperatures due to higher effective charge-carrier mobility and steeper subthreshold swing (enabling lower V_T at fixed I_{OFF}) at lower temperatures [24]. The drive current gain at cryogenic temperatures can be traded for a lower operating voltage (V_{DD}) to improve energy efficiency.

With continued MOSFET miniaturization following Moore's Law, the issue of defect-induced noise, also known as random telegraph noise (RTN), becomes increasingly significant [25, 26]. For instance, silent data corruption, an erroneous computational circuit operation that often goes undetected by error reporting mechanisms, can manifest as application-level problems [27]. In the context of preparing for cryogenic digital computing, it is urgent to explore how RTN in short-channel MOSFETs is affected by low-temperature operation. Equally important is determining whether RTN will be alleviated in advanced MOSFET structures, such as gate-all-around FETs (GAAFETs) being adopted for sub-3nm CMOS technology nodes.

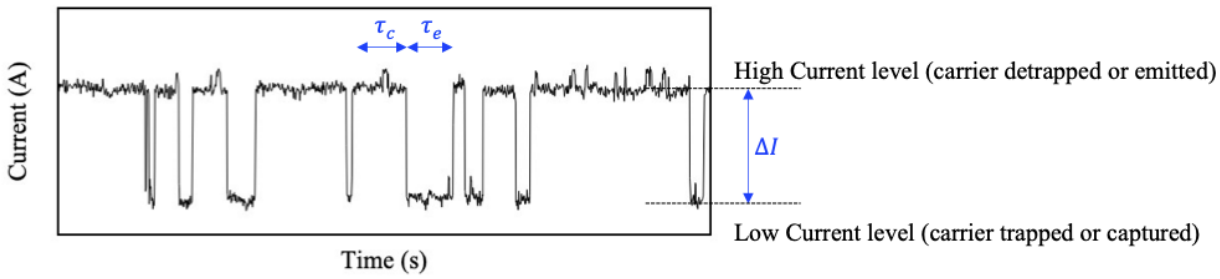
This chapter covers defect-induced noise in MOSFETs as a function of temperature, down to 77K, across different transistor structures. FinFETs fabricated using Samsung's 14nm process technology [28] are used to study the dependence of defect-induced noise on temperature and applied voltages (bias conditions). Unlike prior works that focused on RTN for MOSFET operation in the linear regime, with applied drain voltage below 50 mV [29, 30, 31, 32], this work investigates RTN in the saturation region of operation, with drain voltages ranging from 0.4 V to 0.8 V. Additionally, technology computer-aided design

(TCAD) simulation using Sentaurus (Synopsys, Inc.) [33] is employed to provide insight into the RTN phenomenon and to compare RTN characteristics in advanced MOSFET designs.

2.2 RTN Theory

In the on state, with fixed applied voltages, a MOSFET exhibits drive current that fluctuates over time. This fluctuation is caused by trapping and detrapping of charge carriers from and to the inversion-layer channel due to defects that give rise to locally allowed energy states for electrons and holes within the gate oxide. When a charge carrier is captured by a trap, the transistor current drops; and when the charge carrier is emitted back into the inversion layer, the transistor current returns to its high level. The capture and emission processes occur repeatedly, resulting in current fluctuations over time under constant applied voltages.

Three RTN parameters are defined from the time-dependent current fluctuation shown in Figure 2.1: current reduction (ΔI), capture time (τ_C), and emission time (τ_E).



- ΔI : Current degradation from trapping
- τ_c : mean capture time ($\frac{1}{\tau_c}$ is capture rate)
- τ_e : mean emission time ($\frac{1}{\tau_e}$ is emission rate)

Figure 2.1: Parameters defining RTN characteristics.

ΔI is the amount of current degradation induced by carrier trapping. Two main theories explain this degradation: the number fluctuation theory and mobility fluctuation theory. The number fluctuation theory attributes the current reduction to a decrease in the channel carrier density caused by trapping events. On the other hand, the mobility fluctuation theory explains the degradation as a result of the columbic scattering induced by trapped charges in the gate oxide. In most cases, the observed current degradation is described by a

combination of both number and mobility fluctuation mechanisms.

τ_C, and τ_E are the characteristic times required for carriers to be captured by and emitted from a trap. Since the capture and emission events follow a Poisson process with constant transition rates, both τ_C and τ_E follow exponential distributions. Therefore, the mean values of τ_C and τ_E extracted from time-domain measurements should agree with those obtained by exponential fitting. Figure 2.2 confirms that the mean values and the exponentially fit values extracted from the same measurement data exhibit good agreement for both capture and emission. In this study, all τ_C, and τ_E values are determined from exponential fitting parameters, ensuring a sufficient number of events for reliable analysis.

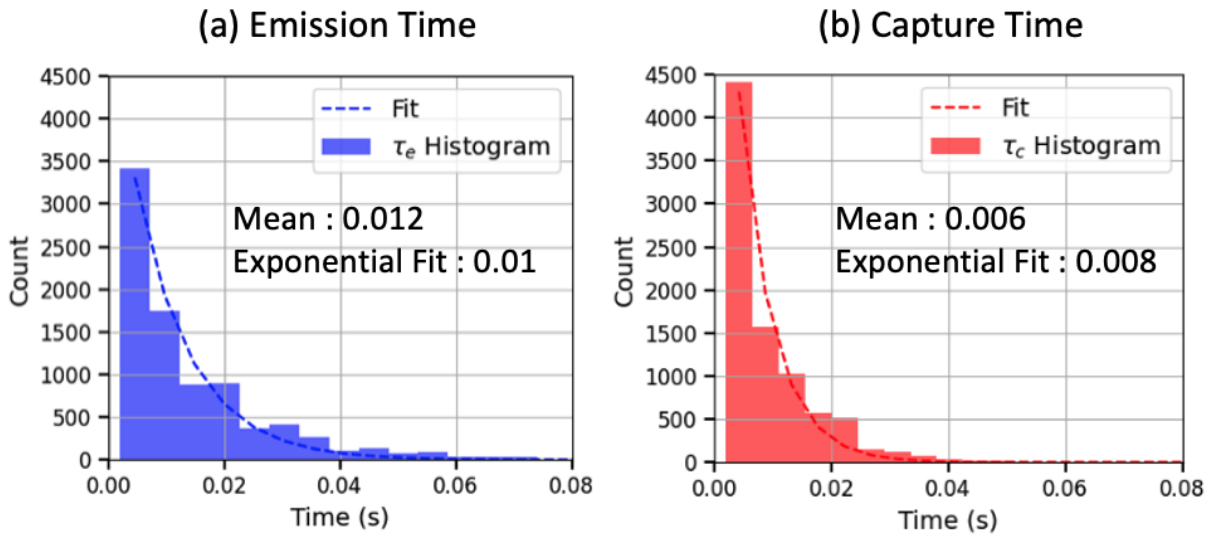


Figure 2.2: Comparison of (a) emission time (τ_E), and (b) capture time (τ_C) extracted from mean values and exponential fitting.

Charge carrier capture and emission rates, which are inverses of τ_C, and τ_E, can be modeled using Shockley-Read-Hall (SRH) recombination theory as follows [34]:

- Capture Rate $= \sigma v n$
- Emission Rate $= \frac{\sigma v n}{g} \exp\left(\frac{E_t - E_f}{kT}\right)$

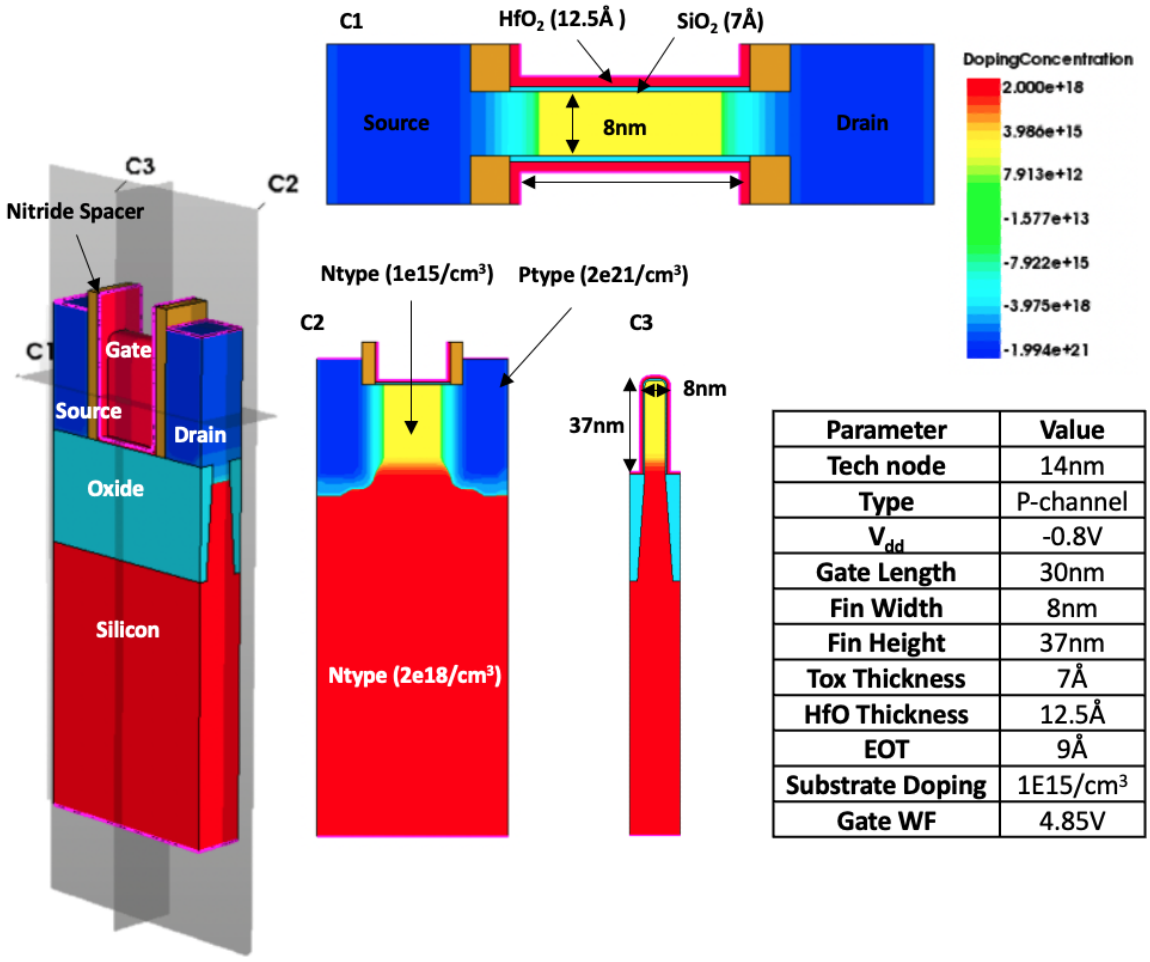
where σ and v are the defect/trap capture cross-section and carrier velocity, respectively, and n is the carrier density in the inversion-layer channel. g is a degeneracy factor. The capture

rate becomes faster as the carrier velocity and carrier density increase. For the emission rate, the alignment between trap energy level in the gate oxide layer and Fermi energy level in silicon should be additionally considered. When the energy level difference between trap level and Fermi level exceeds the thermal energy (kT), carriers cannot overcome the barrier, thereby suppressing the emission process.

2.3 RTN Measurement Results

Samsung 14 nm FinFET Device Structure

Figure 2.3 shows a schematic illustration of Samsung's 14 nm FinFET technology. The device dimensions were adapted from [28]. The gate length (L_G) is 30 nm, the fin width (W_{fin}) and fin height (H_{fin}) are 8 nm and 37 nm, respectively. The interfacial oxide and HfO_2 thicknesses are 7 Å and 12.5 Å, respectively, resulting in a total equivalent oxide thickness (EOT) of 9 Å. The substrate doping concentration is 1×10^{15} cm^{-3}, and the device consists of two electrically connected fins. All RTN measurements in this study were performed in the strong inversion region with $|V_G| = 0.4$ V - 0.8 V, where the threshold voltage is $|V_T| = 0.25$ V.



Parameter	Value
Tech node	14nm
Type	P-channel
V_{dd}	-0.8V
Gate Length	30nm
Fin Width	8nm
Fin Height	37nm
Tox Thickness	7Å
HfO Thickness	12.5Å
EOT	9Å
Substrate Doping	1E15/cm^3
Gate WF	4.85V

Figure 2.3: Samsung's 14nm p-channel FinFET structure and design parameter values used in this work.

RTN Measurement Setup

RTN measurements were taken using a Keysight B1500A semiconductor device parameter analyzer. Specifically, the transistor current was sampled at intervals of 2 ms (corresponding to a frequency of 500 Hz), generating 100,000 data points per measurement instance. A Lakeshore cryogenic probe station was utilized to probe the device under test, with cryogenic conditions achieved by cooling the temperature down to 77 K using liquid nitrogen.

Figure 2.4 presents RTN measurements taken at 300 K, 180 K, and 77 K for a p-channel FinFET. The data confirms the presence of a single trap in the device, exhibiting two discrete current levels. The number of current levels (N_{levels}) depends on the number of traps (n), as follows:

- $N_{\text{levels}} = 2^n$

All RTN data presented in in this chapter were obtained from the same transistor and originate from the same individual trap.

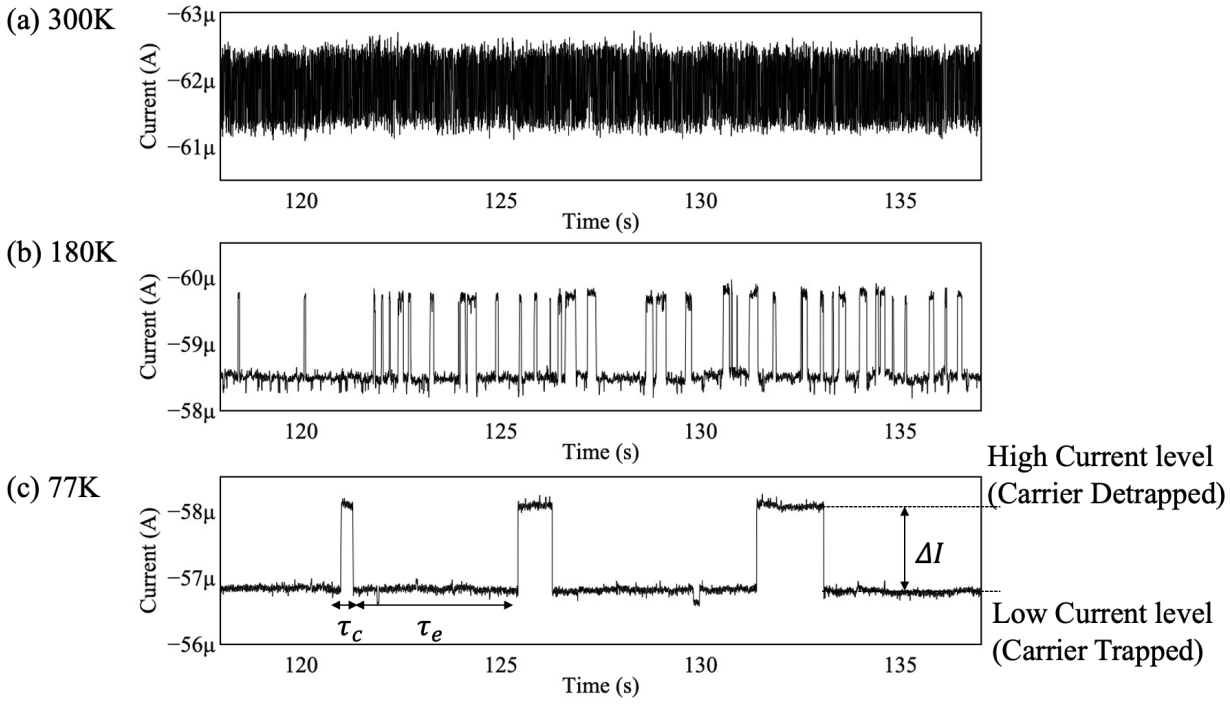


Figure 2.4: p-channel FinFET RTN measurements taken at (a) 300 K, (b) 180 K, and (c) 77 K with $V_G = V_D$ = -0.6 V.

RTN Comparison between n-channel and p-channel FinFETs

The magnitude (ΔI) of RTN in Samsung's 14nm FinFET is greater for p-channel FinFETs than for n-channel FinFETs, as can be seen from Figure 2.5.

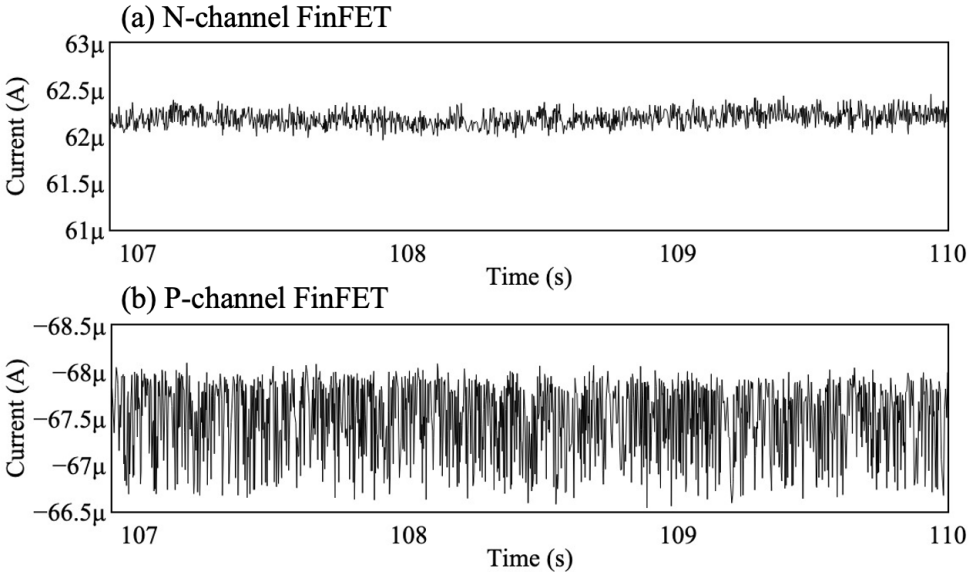


Figure 2.5: RTN measurements taken at 300 K for (a) n-channel and (b) p-channel FinFETs with $|V_G| = |V_D| = 0.6$ V

According to [35], there are two types of RTN mechanisms: one is carrier exchange between the silicon channel (Trap Type1) and gate oxide and the other is carrier exchange between the metal gate and gate oxide (Trap Type2); these mechanisms can be distinguished by the V_G dependence of capture and emission rates. If the trap is located near the channel interface, exchanging carriers between channel and gate oxide (Trap Type1 in Figure 2.6), capture rate increases and emission rate decreases (τ_C/τ_E decreases) with increasing $|V_G|$. This behavior is attributed to the trap level shift below the Fermi level with higher $|V_G|$, making carrier capture more favorable and emission less likely. On the other hand, if the trap is located near the metal gate, exchanging carriers between metal and gate oxide (Trap Type2 in Figure 2.6), the capture rate decreases and emission rate increases (τ_C/τ_E increases) with increasing $|V_G|$. In this case, the metal work function shifts with increasing $|V_G|$ to suppress the capture process and enhance emission.

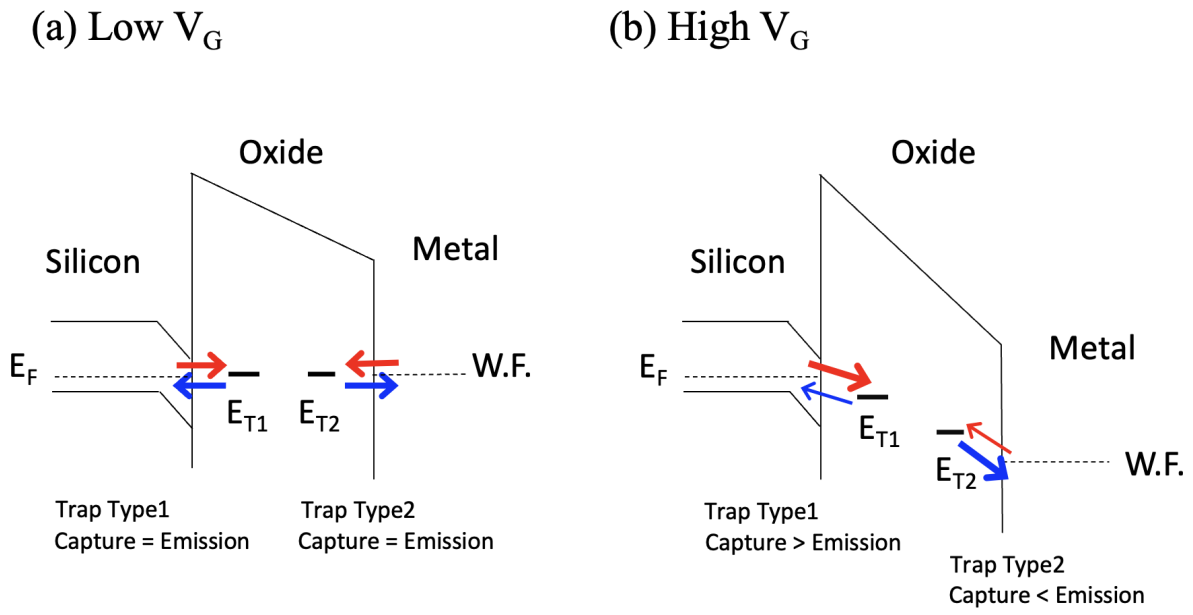


Figure 2.6: Schematic illustration of electron capture and emission behavior in an n-channel MOSFET at (a) low V_G and (b) high V_G for two types of oxide trap: near silicon and oxide interface (Trap Type1) and near metal and oxide interface (Trap Trap2)

Figure 2.7 shows that Samsung's 14nm n-channel and p-channel FinFETs exhibit distinctly different behaviors with increasing $|V_G|$.

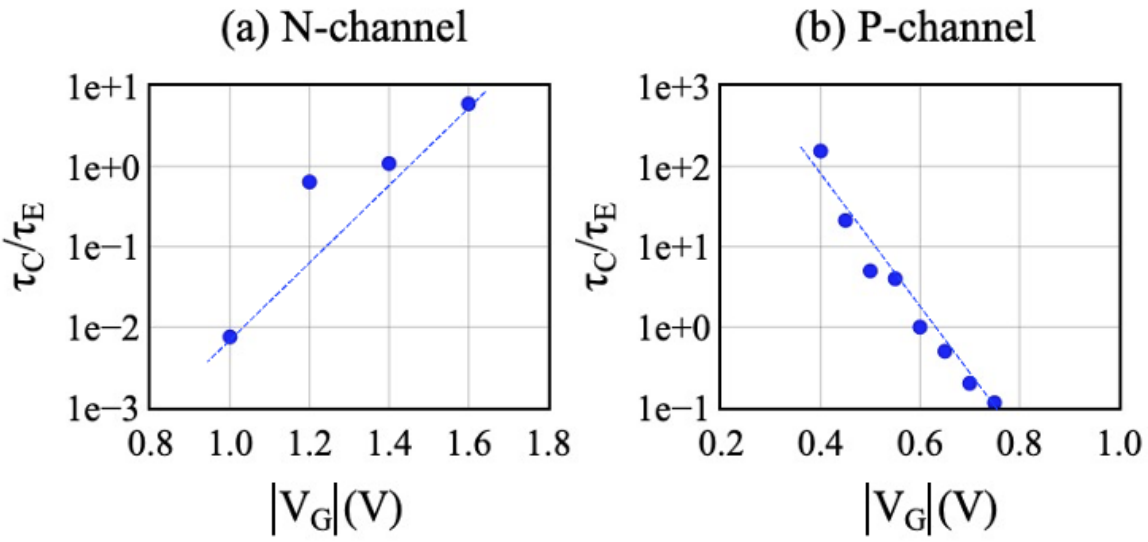


Figure 2.7: τ_C/τ_E as a function of $|V_G|$ for (a) n-channel and (b) p-channel FinFETs.

For the p-channel FinFET, τ_C/τ_E decreases with increasing $|V_G|$, indicating silicon channel side trapping, whereas for the n-channel FinFET, τ_C/τ_E increases with increasing $|V_G|$, indicating metal side trapping. Since traps located near the gate side have a weaker impact on inversion-layer carrier mobility, ΔI is smaller for the n-channel FinFET. Therefore, this study is focused on RTN in p-channel FinFETs and GAAFETs.

ΔI Analysis

Figure 2.8 shows how ΔI depends on the applied gate voltage (V_G) and drain voltage (V_D), respectively, at various temperatures.

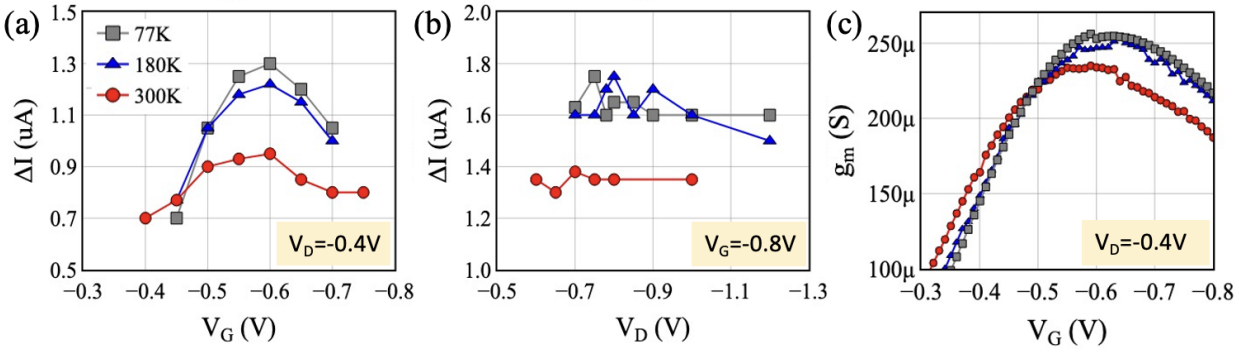


Figure 2.8: Measured ΔI for a p-channel FinFET at various temperatures as a function of (a) V_G and (b) V_D. (c) Measured g_m of a p-channel FinFET as a function of V_G, at various temperatures.

It should be noted that V_T does not vary much with temperature, as can seen from 2.8 (c) and Figure 2.14, so that comparisons can be made for the same value of V_G. ΔI increases with decreasing temperature as Coulombic scattering becomes increasingly dominant, causing a trapped charge to degrade inversion-layer carrier mobility more significantly. The non-monotonic relationship between ΔI and V_G—where ΔI increases asV_G increases in magnitude until V_G = -0.6 V, beyond which it begins to decrease—can be explained by the dependence of the transconductance (g_m) on V_G (Figure 2.8 (c)) which reflects the degradation of effective carrier mobility for large transverse electric field. V_D has a relatively weak influence on ΔI because it does not significantly alter the inversion-layer carrier mobility.

Capture and Emission Rates Analysis

As $|V_G|$ increases, the hole capture rate increases due to increasing hole density (p) with increasing gate overdrive voltage ($|V_T|$ = 0.25 V). Conversely, the emission rate decreases as $|V_G|$ increases (Figure 2.9 (a)). It is because the trap energy level (E_t) moves further away from the Fermi energy level (E_f) in the inversion-layer channel as depicted in Figure 2.9 (b), suppressing hole emission from the trap.

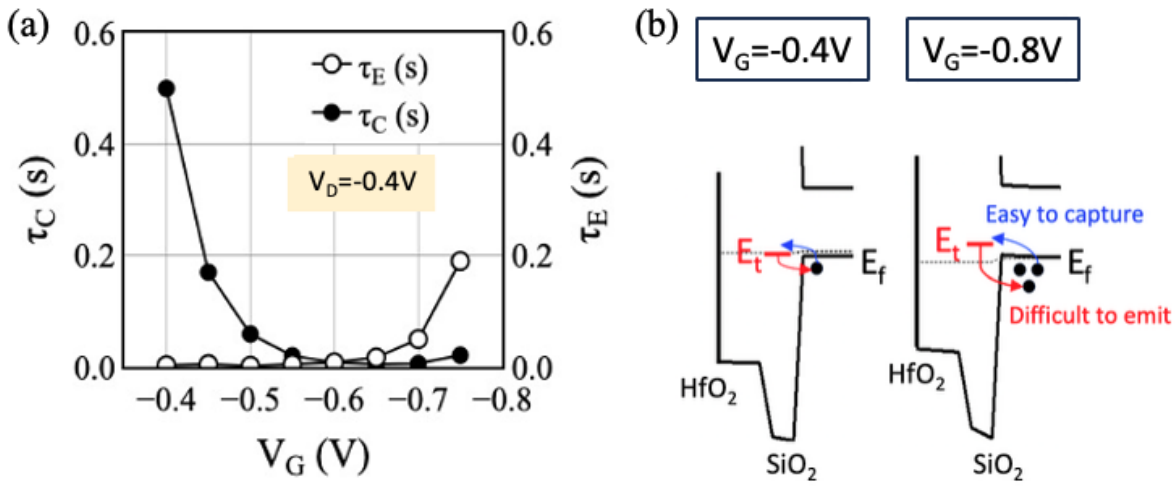


Figure 2.9: (a) Measured mean capture and emission times of a p-channel FinFET as a function of V_G (T = 300 K). (b) Energy-band diagrams illustrating increasing capture rate and decreasing emission rate with increasing $|V_G|$ in a p-channel MOSFET.

Next, the mean values of measured capture and emission times are plotted in Figure 2.10 (a) as a function of temperature. Typically, hole velocity (v_h) is approximated as thermal velocity (v_{th}). Due to reduced thermal velocity (v_{th}) at lower temperatures, it takes longer for a hole to be trapped or detrapped.

Interestingly, capture and emission times were experimentally observed to decrease significantly with increasing V_D in the saturation region of operation (Figure 2.10 (b)). This trend has not been widely reported in the literature.

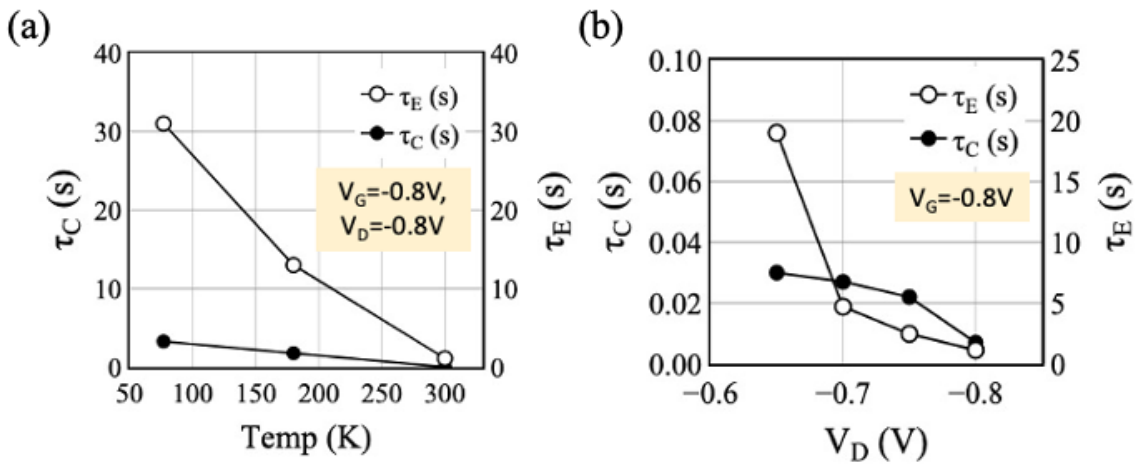


Figure 2.10: Measured mean capture and emission times of a p-channel FinFET as a function of (a) temperature, and (b) V_D (T = 300 K).

A previous study [36] also reports a high dependence of capture and emission rates on V_D in the saturation region of operation. Considering that higher V_D reduces gate control, it would be expected that capture rate deceases while emission rate increases with increasing V_D. However, in the present study, both capture and emission rates increase with increasing V_D at different temperatures (Figure 2.11), suggesting that a different underlying mechanism is responsible. To describe this V_D trend, the formula for v_h—previously approximated as v_{th}—should be reexamined. In nanometer-scale MOSFETs operating with $|V_D|$ greater than 50 mV, drift velocity (v_d) can become comparable to thermal velocity. Moreover, it was argued in [37] that carrier temperature (T_h) should be incorporated into the carrier velocity formula because it provides thermal energy for activating traps. Taking these factors into account, the carrier velocity formula is:

$$v_h = \sqrt{v_{\text{th}}^2 \frac{T_h}{T} + v_d^2}$$

From this formula, it can be deduced that an increase in $|V_D|$ should increase v_h (and hence the emission rate) due to increased drift velocity, and that this dependence on $|V_D|$ should be stronger at lower temperatures due to lower v_{th}. The experimental results in Figure 2.11 confirm a more dramatic change in capture and emission rates at lower temperatures.

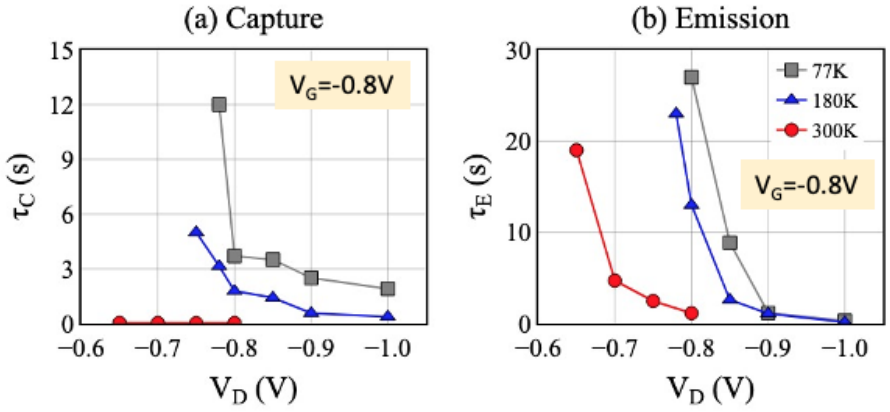


Figure 2.11: Measured (a) mean capture time and (b) emission time vs. V_D for a p-channel FinFET at various temperatures.

RTN Power Spectral Density

The power spectral density (PSD) of a single RTN trap is described by a Lorentzian function,

$$S_I(f) = \frac{4(\Delta I)^2 \tau}{1 + (2\pi f \tau)^2},$$

where ΔI is the current fluctuation amplitude and τ is the effective time constant given by

$$\tau = \frac{\tau_C \tau_E}{\tau_C + \tau_E}.$$

At low frequencies ($f \ll f_c$), the PSD approaches a constant value proportional to $(\Delta I)^2$, whereas at high frequencies ($f \gg f_c$), it decreases following a $1/f^2$ dependence. The corner frequency f_c, which separates these two regimes, is given by

$$f_c = \frac{1}{2\pi\tau} = \frac{1}{2\pi}\left(\frac{1}{\tau_C} + \frac{1}{\tau_E}\right).$$

Figure 2.12 plots RTN power spectral density measured at 300 K, 180 K, and 77 K. It is observed that f_c decreases at lower temperatures due to reduced capture and emission rates. Moreover, at low frequency range ($f \ll f_c$), PSD is higher at 77 K than at 300 K due to larger ΔI. For analog circuit applications, it is desirable to minimize ΔI to fully benefit from the lower capture and emission rates at cryogenic temperatures.

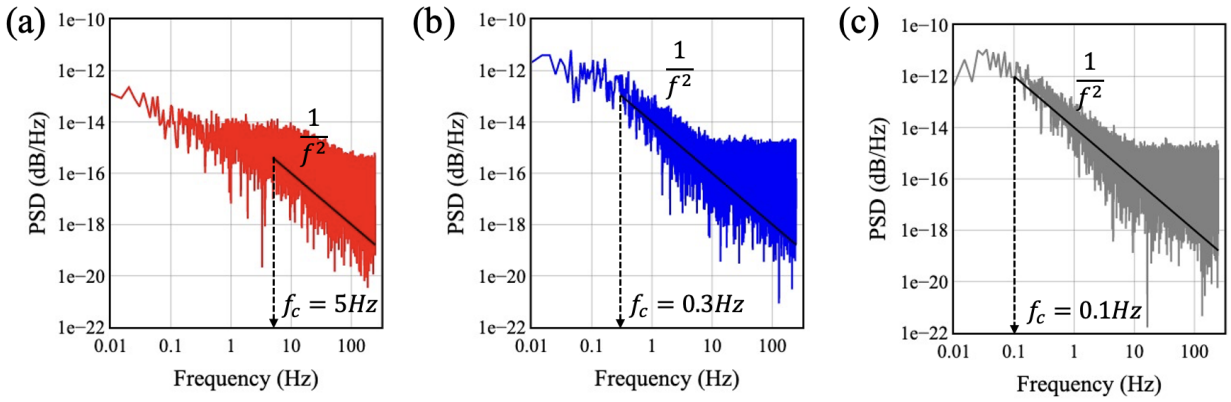


Figure 2.12: RTN power spectral density measured at (a) 300 K, (b) 180 K, and (c) 77 K for a p-channel FinFET with $V_G = V_D$ = -0.8 V.

2.4 RTN Modeling Results

The RTN modeling using TCAD simulation followed the procedure illustrated in Figure 2.13.

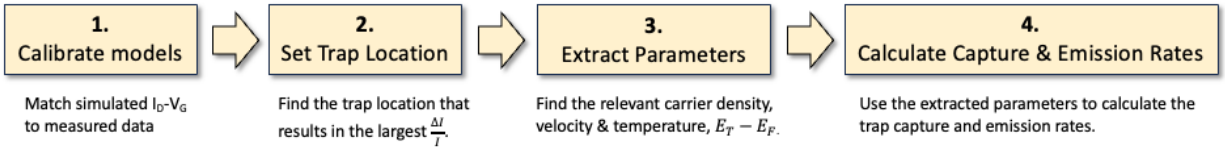


Figure 2.13: RTN simulation and modeling procedure.

First, the model parameters were calibrated so that the I_D–V_G characteristics for the simulated structure (Figure 2.3) match the measured data, as shown in Figure 2.14.

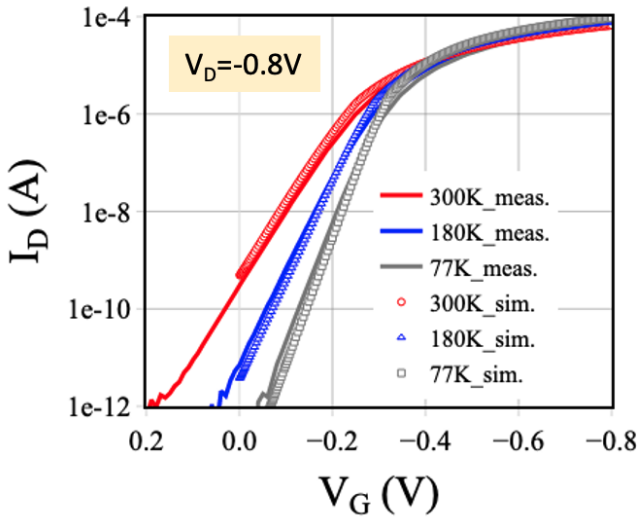


Figure 2.14: Measured and simulated I_D–V_G characteristics of p-channel FinFET.

Further simulations are performed to identify the physical trap location that results in the largest $\Delta I/I$, which corresponds to the highest current density (Figure 2.15).

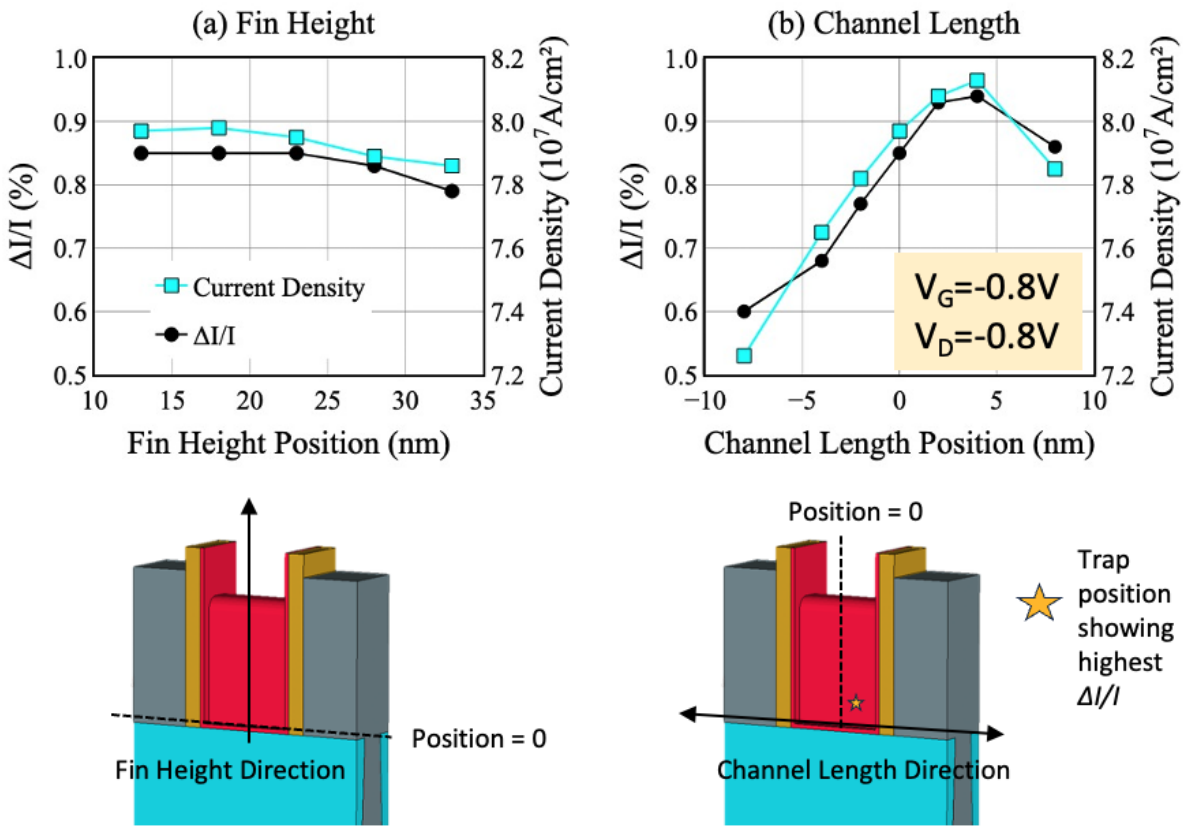


Figure 2.15: Simulated current degradation ($\Delta I/I$) and current density of p-channel FinFET at 300 K with changing trap position in the (a) fin height and (b) channel length direction.

Parameter values for calculating the capture and emission rates are then extracted based on the trap location and applied to the modified Shockley–Read–Hall (SRH) equation. Figure 2.16 compares the modeled and measured values of capture and emission rates as functions of V_G, V_D, and temperature, respectively, showing good agreement between the RTN model and measurements.

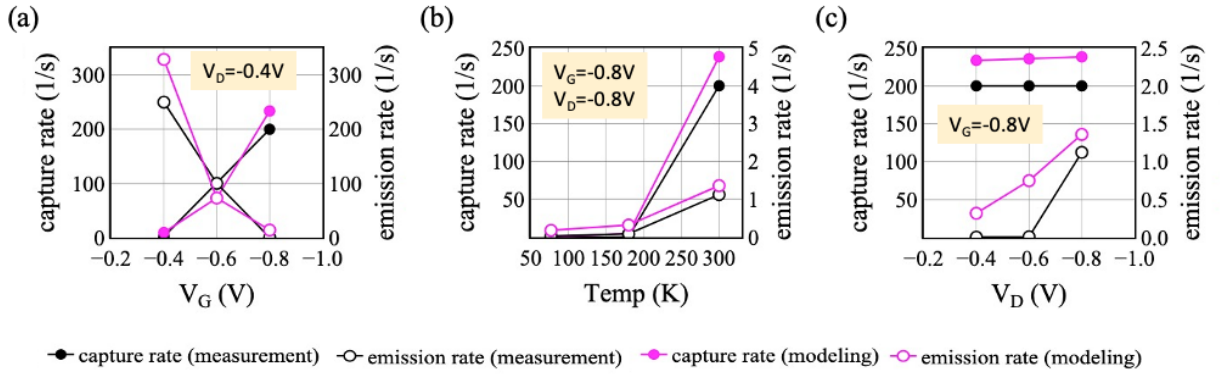


Figure 2.16: Modeling results illustrating capture and emission rates as a function of (a) V_G (T = 300 K), (b) temperature, and (c) V_D (T = 300 K).

2.5 Comparison of RTN in Advanced MOSFETs

TCAD simulation using models calibrated with measured data was employed to analyze RTN in advanced MOSFETs: the 3nm-generation FinFET, 2nm-generation GAAFET with 30 nm nanosheet width ("Wide GAAFET"), and 2nm-generation GAAFET with 10 nm nanosheet width ("Narrow GAAFET") (Figure 2.17) [38].

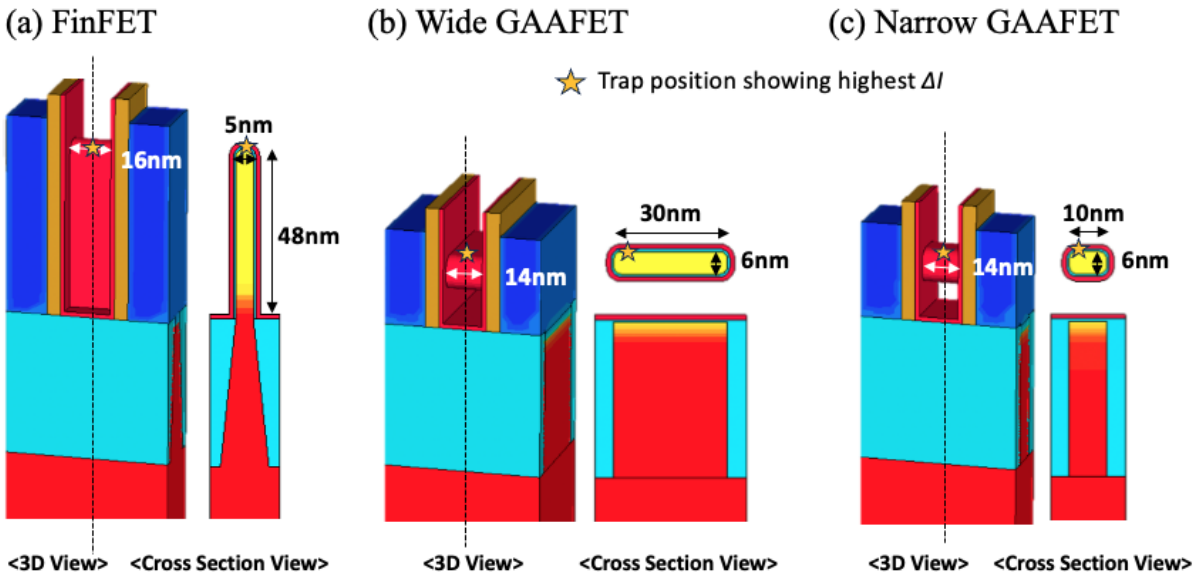


Figure 2.17: Simulated hole current density profiles showing the influence of a trapped hole at the location marked with a star. T = 77 K, $V_G = V_D$ = -0.6 V.

Table 2.1 compares the simulated performance and RTN characteristics for the different transistor designs and operating conditions. The Wide GAAFET operating at V_{DD} = -0.65 V has almost identical on-state current as the FinFET operating at V_{DD} = -0.7 V, while the Narrow GAAFET operating at V_{DD} = -0.65 V has the same subthreshold swing as the FinFET operating at V_{DD} = -0.7 V. V_{DD} was reduced to -0.6 V in comparing the three transistors for 77 K operation.

Table 2.1: Design parameters and operating conditions for advanced p-channel MOSFETs, along with their simulated performance and RTN characteristics.

Splits		FinFET		Wide GAAFET		Narrow GAAFET	
Design Parameter	EOT (A)	9		9		9	
	Gate Length (nm)	16		14		14	
	Fin Width (nm)	5					
	Fin Height (nm)	48					
	Nanosheet Thickness (nm)			6		6	
	Nanosheet Width (nm)			30		10	
	Total Width per Fin/Nanosheet (nm)	101		72		32	
	Number of Fin/Nanosheet	2		3		3	
	Total Width (nm)	202		216		96	
	Work Function (eV)	4.85		4.816		4.874	
Operating Condition	Operating Temperature (K)	300	77	300	77	300	77
	Operating Voltage (V_{DD}=V_G=V_D)	-0.7	-0.6	-0.65	-0.6	-0.65	-0.6
Id-Vg Characteristic	On-state Current per Fin/Nanosheet (A)	8.10E-05	7.80E-05	5.16E-05	5.32E-05	2.53E-05	2.68E-05
	Off-state Current per Fin/Nanosheet (A)	5.13E-10	1.99E-14	3.54E-10	1.35E-14	3.33E-10	1.47E-14
	Total On-state Current (A)	1.62E-04	1.56E-04	1.55E-04	1.60E-04	7.60E-05	8.04E-05
	Total Off-state Current (A)	1.03E-09	3.97E-14	1.06E-09	4.05E-14	9.98E-10	4.41E-14
	Threshold Voltage (V)	-0.20	-0.29	-0.21	-0.31	-0.20	-0.29
	Subthreshold Swing (mV/dec)	64	37	67	40	64	37
RTN Characteristic	ΔI (A)	1.08E-06	1.61E-06	1.11E-06	1.53E-06	1.06E-06	1.66E-06
	$\Delta I/I$ (%)	0.67	1.03	0.72	0.96	1.40	2.10
	Capture Rate (1/s)	289	0.39	243	0.41	406	0.72
	Emission Rate (1/s)	44	2.5E-05	31	9.6E-06	21	4.1E-06

Consistent with the measured RTN trends, the simulations show a larger ΔI at lower temperature. At 77 K, the Narrow GAAFET exhibits the largest ΔI, whereas the Wide GAAFET exhibits the smallest ΔI. Since the nanosheet thickness (6 nm) is greater than the fin width (5 nm), the influence of trapped charge is less extensive in the Wide GAAFET than in the FinFET. However, the trapped charge influences the largest fraction of the channel region for the Narrow GAAFET (Figure 2.18).

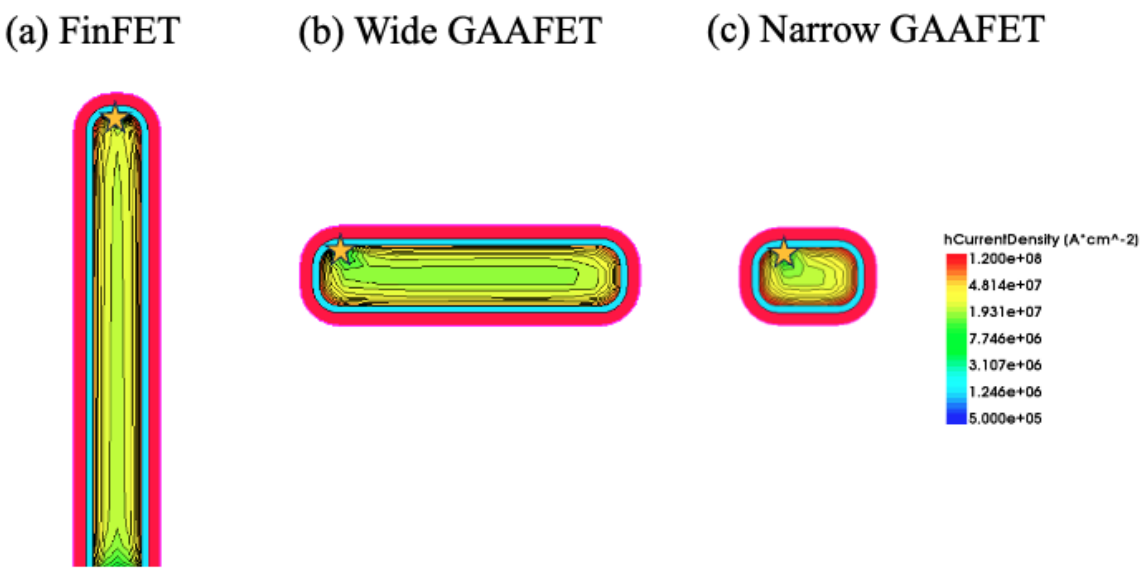


Figure 2.18: Simulated hole current density profiles showing the influence of a trapped hole at the location marked with a star. T = 77K, $V_G = V_D$ = -0.6 V.

Calculated capture and emission rates in the advanced MOSFET structures are much lower at cryogenic temperature than at room temperature, particularly for the emission rate. This is because the capture rate only reflects the reduction in carrier velocity as temperature decreases, whereas the emission rate is influenced not only by carrier velocity but also by the thermal energy (kT) that enables carriers to overcome the energy barrier between the trap level and Fermi level.

2.6 Conclusion

The impact of RTN becomes more significant in scaled devices. Therefore, it is important to predict RTN behavior in advanced technology nodes such as FinFETs and GAAFETs. This study provides quantitative analysis of the current fluctuation amplitude (ΔI) as well as the capture and emission rates in Samsung's 14nm FinFET technology.

A new trend in the dependence of capture and emission rates on V_D is identified, which is attributed to the carrier drift velocity becoming comparable to the thermal velocity in short-channel devices. Based on this observation, a modified carrier velocity model that incorporates both thermal and drift components is proposed.

Using this model, RTN behavior is predicted at cryogenic temperatures in advanced transistors. Cryogenic operation of MOSFETs will reduce carrier capture and emission rates and hence improve low frequency noise characteristics for analog circuit applications. To mitigate the increase in RTN magnitude (ΔI) at low temperatures, the fin/nanosheet channel region thickness and width should be increased within other design constraints.

Chapter 3

Effects of aluminum in the metallic gate of metal-oxide-semiconductor (MOS) devices with HfO_2 dielectric

3.1 Introduction

Over the past two decades, the integration of high-permittivity (high-k) oxide and metallic gate (HKMG) films into the gate stack of complementary metal-oxide-semiconductor field-effect transistors (CMOSFETs) has provided for improved electrostatic integrity while maintaining low gate leakage current, facilitating continued transistor performance and higher inversion charge. Careful engineering of the metallic gate film is crucial for advanced CMOS technologies employing thin-body transistor structures such as fin-shaped semiconductor and "gate-all-around" (GAA) FETs, since their threshold voltage (V_T) is tuned by adjusting the effective gate work function. Although details of the most advanced metallic gate films have not been publicly disclosed by leading semiconductor manufacturers, multiple publications [39, 40, 41] indicate that a titanium-aluminum (TiAl) alloy layer on top of a thin titanium nitride (TiN) barrier layer is used to achieve the low gate work function required for n-channel (NMOS) transistors. However, the mechanism by which aluminum (Al) affects the effective gate work function, as well as tradeoffs associated with incorporating Al into the metallic gate film, need further exploration.

This chapter presents a detailed study to elucidate the impact of intermixing at the metallic gate/high-k oxide interface in a MOS system comprising a thin HfO_2 dielectric film (2.5–8.0 nm thick). The metallic gate film comprises an ultrathin Al layer (1–3 nm thick) sandwiched between TiN layers formed by atomic layer deposition (ALD). Unlike previous studies that focused on homogeneous TiAl alloy gate electrodes, this work investigates an ultrathin Al insertion layer embedded within a TiN gate stack. This multilayer gate structure more closely resembles the practical HKMG stacks employed in advanced CMOS

technologies.

Using this gate structure, it is first shown that Al in the metal gate film results in electron trap states in the HfO_2 film near the metal gate, as revealed by capacitance–voltage (C–V) measurements.

Next, we investigated the effect of the ultrathin Al layer thickness on dielectric interface properties, by tracking changes in flat-band voltage (V_{FB}) from C–V measurements and changes in electron energy barrier height from I–V measurements. Elemental depth profiling and chemical state analyses of Al and Ti were performed using electron-dispersive spectroscopy (EDS) and electron energy loss spectroscopy (EELS) in conjunction with high-resolution cross-sectional transmission electron microscopy (HR-XTEM) of MOS capacitor samples. The results show that the effective gate work function monotonically reduces with increasing ultrathin Al layer thickness, due to an oxygen scavenging effect that reduces the oxygen content in TiON. Furthermore, a forming gas anneal (FGA) at 350 °C results in tradeoffs of increased gate leakage and degraded dielectric breakdown voltage due to oxygen redistribution out of the HfO_2 film.

Lastly, we examined dielectric degradation for MOS devices with Al gate. It is found that a TiN diffusion barrier layer is essential to prevent upward diffusion of Si and Hf into the Al gate, and thereby to mitigate increased gate leakage.

3.2 Experimental

Long-channel MOSFETs were fabricated on p-type silicon wafer substrates doped with 1.0×10^{15} cm^{-3} boron, using a non-self-aligned source/drain (S/D) process and p-n junctions for device isolation. Masked ion implantation of phosphorus (30 keV, 2.0×10^{15} cm^{-2} + 15 keV, 2.0×10^{15} cm^{-2}) and boron difluoride (30 keV, 4.0×10^{15} cm^{-2}) was used to form heavily doped n-type S/D and p-type body contact regions, respectively, through a screening oxide layer (10 nm thick). Dopant activation was achieved via rapid thermal anneal (RTA) at 1000 °C for 5 sec in N_2 ambient. Afterwards, the screening oxide layer was removed using a dilute HF solution (25:1 H_2O:49 % HF by volume). Immediately following oxide removal, 8 Å chemical oxide was grown by immersing the substrate in a heated solution of H_2O:NH_4OH:H_2O_2 (5:1:1) by volume. HfO_2 was subsequently deposited by atomic layer deposition (ALD) at 250 °C using tetrakis(dimethylamido)hafnium(IV) (TDMAHf, $Hf(NMe_2)_4$) as the Hf precursor and H_2O as the oxidant. Prior to the ALD cycles, 5 cycles of H_2O pulses were applied to improve the quality of the chemical oxide. For HfO_2 thicknesses of 2.5, 3.5, and 8.0 nm, the deposition employed 29, 40, and 90 ALD cycles, respectively, with each cycle consisting of sequential TDMAHf and H_2O pulses. HfO_2 film thickness was measured using spectroscopic ellipsometry. To form the gate film, a 1.5 nm-thick TiN barrier layer was deposited by plasma-enhanced ALD at 300 °C using tetrakis-dimethylamido-titanium (TDMAT; $Ti(NMe_2)_4$) as the Ti precursor and H_2/N_2 plasma as the reactant, with 12 ALD cycles. Then, an ultrathin Al layer (1-3 nm thick) was deposited by electron-beam

evaporation with an in-situ crystal monitor to measure the deposited layer thickness. Afterwards, another TiN layer (14 nm thick) was deposited by ALD, followed by a capping layer stack consisting of 6 nm TiN and 40 nm Al-Si (2 % silicon) deposited using a 5 kHz pulsed-DC sputter system. It should be noted that the surface of the ultrathin Al layer is naturally oxidized upon exposure to air as the wafer is transferred from the evaporator into the ALD tool. The metallic gate layer stack was patterned using lithography and reactive ion etching (RIE), stopping on the gate-dielectric HfO_2 layer. Subsequently, a 350 nm-thick SiO_2 interlayer dielectric (ILD) was deposited over the patterned gate electrodes. Then, contact holes were formed in the ILD and residual HfO_2 layer. Afterwards, the contacting metal film (consisting of a 10 nm-thick Ti layer, 10 nm-thick TiN layer, and 40 nm-thick Al-Si layer sequentially deposited using the aforementioned sputtering system) was deposited and patterned. Finally, a forming gas anneal (FGA) was performed in a H_2/N_2 ambient at 350 °C for 30 min. MOS capacitors (MOSCAPs) were fabricated using the same process flow as the MOSFETs without the ion implantation, contact formation, and contacting metal film deposition and patterning steps. The wafer backside was used for MOS capacitor body contact.

Figure 3.1 shows schematic illustrations of the HKMG stack in the MOSFETs (Figure 3.1 (a)), the corresponding top-view layout (Figure 3.1 (b)), and the typical measured electrical characteristics of the fabricated MOSFETs at room temperature (Figures 3.1 (c)–(f)). With an 8 Å chemical oxide interfacial layer (IL), a 2.5 nm ALD-HfO_2 gate dielectric film, and a TiN/1.0 nm Al/TiN gate film, the device exhibits an inversion capacitance of 2.6 μF cm^{-2} and inversion charge (Q_{INV}) of 5.38×10^{-5} C cm^{-2} at $V_G = 1.5$ V, in good agreement with the quantum-mechanical C–V (QMCV) model yielding an equivalent oxide thickness (EOT) of 1.02 nm (Figure 3.1 (c)). The transfer characteristics (I_D–V_G) in Figure 3.1 (d) exhibit a threshold voltage of 0.52 V and a subthreshold swing of 68 mV/dec. Figure 3.1 (e) shows a peak electron mobility of 200 cm^2/V s. These results confirm the reasonable quality of the fabricated HKMG film stack. Figure 3.1 (f) shows the relationship between ALD-HfO_2 thickness and EOT, indicating a dielectric constant of 16.7 for HfO_2 and an EOT of 4.4 Å for the interfacial layer. The thin EOT of the IL suggests that the chemical oxide was converted to hafnium silicate during the device fabrication process. Other possible explanations include the formation of Si sub-oxide states within the interfacial transition region [42] or nitrogen incorporation from TiN layer into the IL.

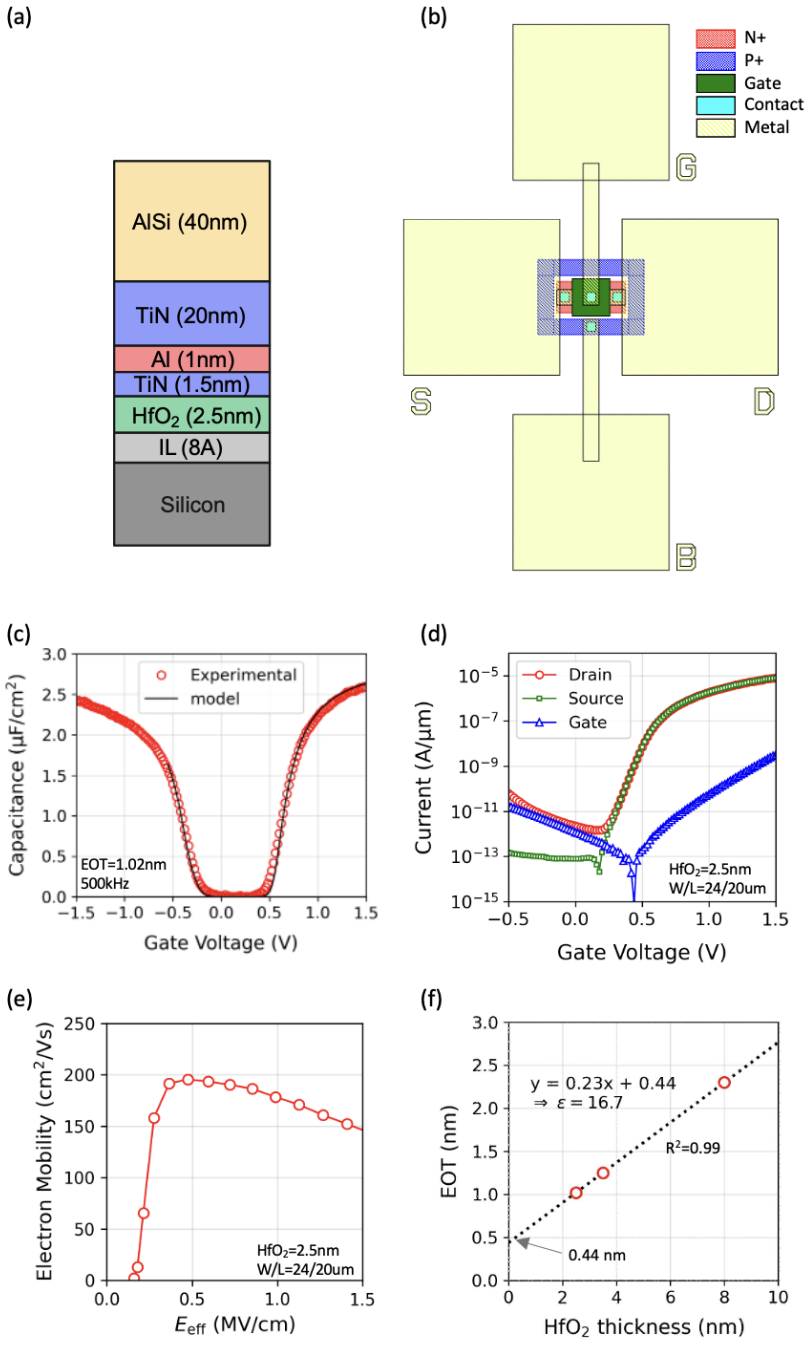


Figure 3.1: (a) Schematic illustration of the HKMG of MOSFETs. (b) Layout of the MOSFETs. Electrical characteristics of MOSFETs with HKMG stack (2.5 nm HfO_2, 1nm ultrathin Al), measured at room temperature: (c) C–V curve compared against QMCV model; (d) transfer characteristics; (e) field-effect electron mobility vs. effective transverse electric field. (f) Equivalent oxide thickness vs. HfO_2 thickness.

3.3 Impact of Al incorporation in the dielectric layer (HfO_2) on interfacial trap states

Figure 3.2 (a) shows typical measured C–V characteristics of the same fabricated MOSFETs, for small-signal frequencies of 1 kHz and 500 kHz. A hump in the 1 kHz C–V curve suggests the presence of electron trap sites within the oxide film.

The frequency dependence of C–V characteristics was further investigated for gate film stacks with vs. without the ultrathin Al layer. Figure 3.2 (b) compares interface trap state density (D_{IT}) vs. energy curves extracted by comparing the ideal QMCV model and the measured C–V curves (inset of Figure 3.2 (c)), for the two gate film stacks. For both, the D_{IT} peak appears at E_T - $E_i = 0.4$ eV which is approximately 0.15 eV below the conduction band of Si, corresponding to the weak inversion region of operation. The extracted trap level is consistent with previous reports of singly charged oxygen vacancies in HfO_2 [43]. Figure 3.2 (c) compares the frequency dependence of the D_{IT} peak value ($D_{IT,max}$) for the two gate film stacks, as a function of small-signal frequency. It can be seen that $D_{IT,max}$ decreases as frequency increases, with the Al-incorporated gate film stack showing higher $D_{IT,max}$ across most of the measured frequency range.

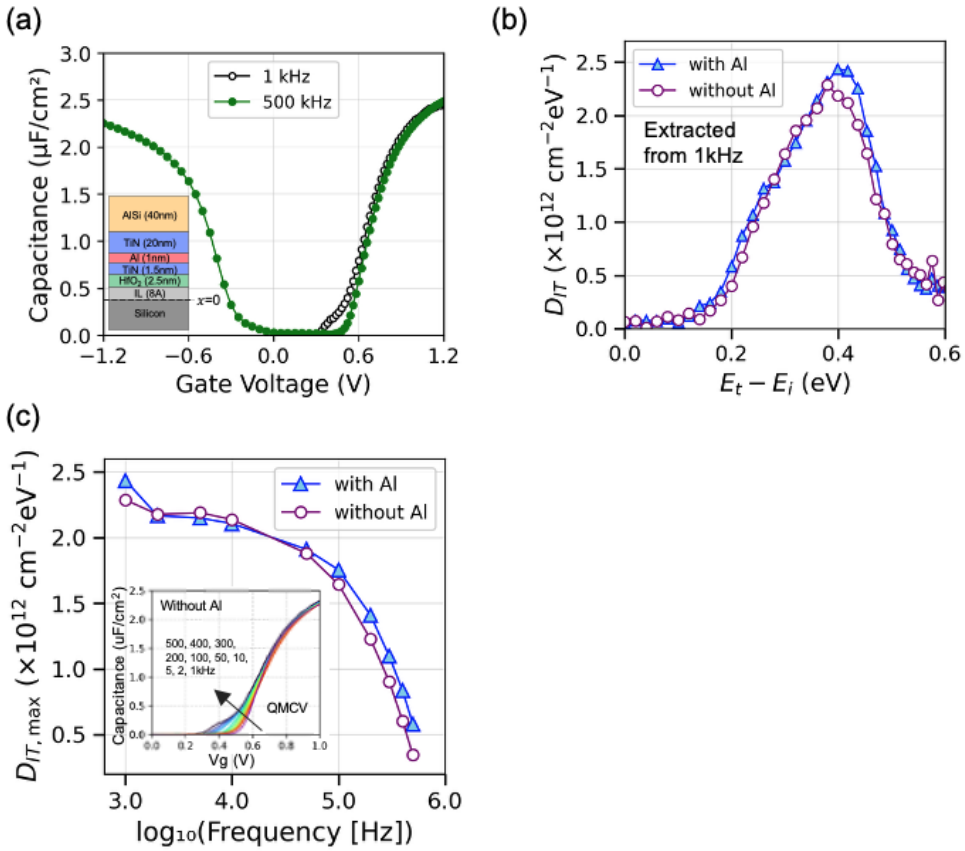


Figure 3.2: Electrical characteristics of MOSFETs with HKMG stack (2.5 nm HfO_2), measured at room temperature: (a) C–V curves for MOSFETs with a 1 nm ultrathin Al layer; (b) interface trap density (D_{IT}) vs. trap energy (E_T - E_i) extracted from 1 kHz C–V curves; and (c) peak D_{IT} vs. C–V measurement frequency for MOSFETs with and without a 1 nm ultrathin Al layer.

The trap density per unit volume, N_T, was calculated assuming that the frequency dependence of $D_{IT,max}$ is due to the time-dependent electron wavefunction penetration depth $x(t)$. According to the tunneling front model [44, 45]:

$$x(t) = \frac{1}{2\beta} \ln\left(\frac{t}{t_0}\right) \tag{3.1}$$

where the decay constant

$$\beta = \sqrt{\frac{2m_{\mathrm{ox}}^{*} q \Phi_{\mathrm{B}}}{\hbar^2}} \tag{3.2}$$

Here, m_{ox}^{*}, q, and Φ_B are the electron effective mass in the dielectric, the elementary charge, and the barrier height at the dielectric/Si interface, respectively, and $t_0 = 10^{-13} \sim 10^{-14}$ s is set by the phonon frequency, representing the electron attempt rate for tunneling through the barrier. Using this model, the time required for penetrating through the 0.8 nm-thick IL (Hf = 15 % silicate based on EOT = 0.44 nm) is approximately 9.3 ps; hence, the traps are located within the HfO_2 film for the entire measurement frequency range studied herein.

N_T was calculated by dividing the difference in $D_{IT,max}$ by the difference in the penetration depth (cf. Eqn. (3.1)) between two frequencies, assuming $m_{ox}^{*} = 0.44$, $\Phi_B = 3.1$ eV for the IL (SiO_2), and $m_{ox}^{*} = 0.17$ [46], $\Phi_B = 1.2$ eV [47] for HfO_2, respectively.

The volume density of trapped electrons for a given period $\tau = 1/2\pi f$ of the C–V measurement, $N_T(x)$, is dependent on the distance from the dielectric interface x:

$$N_{\mathrm{T}}(x) = N_0 e^{-x/\lambda} \tag{3.3}$$

where N_0 is the trap density in the film and λ is the tunneling attenuation length [48, 49]. From Eqns. (3.1) and (3.3), the relationship between N_T and τ is given by:

$$\log_{10} N_{\mathrm{T}}(\tau) = \log_{10}(N_0) - \frac{1}{2\beta\lambda} \log_{10}\left(\frac{\tau}{\tau_0}\right) \tag{3.4}$$

Figure 3.3 (a) plots N_T as a function of τ on a log–log scale for devices fabricated with vs. without the ultrathin Al layer in the metallic gate film stack. For high frequencies (τ below 1 μs), the slopes are similar ($\lambda \sim 0.13$ nm); however, at lower frequencies, the gate film stack with the ultrathin Al layer exhibits a shallower slope, corresponding to a larger tunneling attenuation length ($\lambda \sim 0.16$ nm). This suggests that Al in the metallic gate film stack resulted in an additional electron trap state within the HfO_2. The SIMS depth profile of the metallic gate film stack with the ultrathin Al layer (Figure 3.3 (b)) confirms the presence of Al within the HfO_2 layer near the gate interface. It also shows a smaller Al peak in the HfO_2 near the silicon interface, which likely contributes to interface traps, resulting in the MOSFET subthreshold swing being larger than the ideal value of 60 mV/dec.

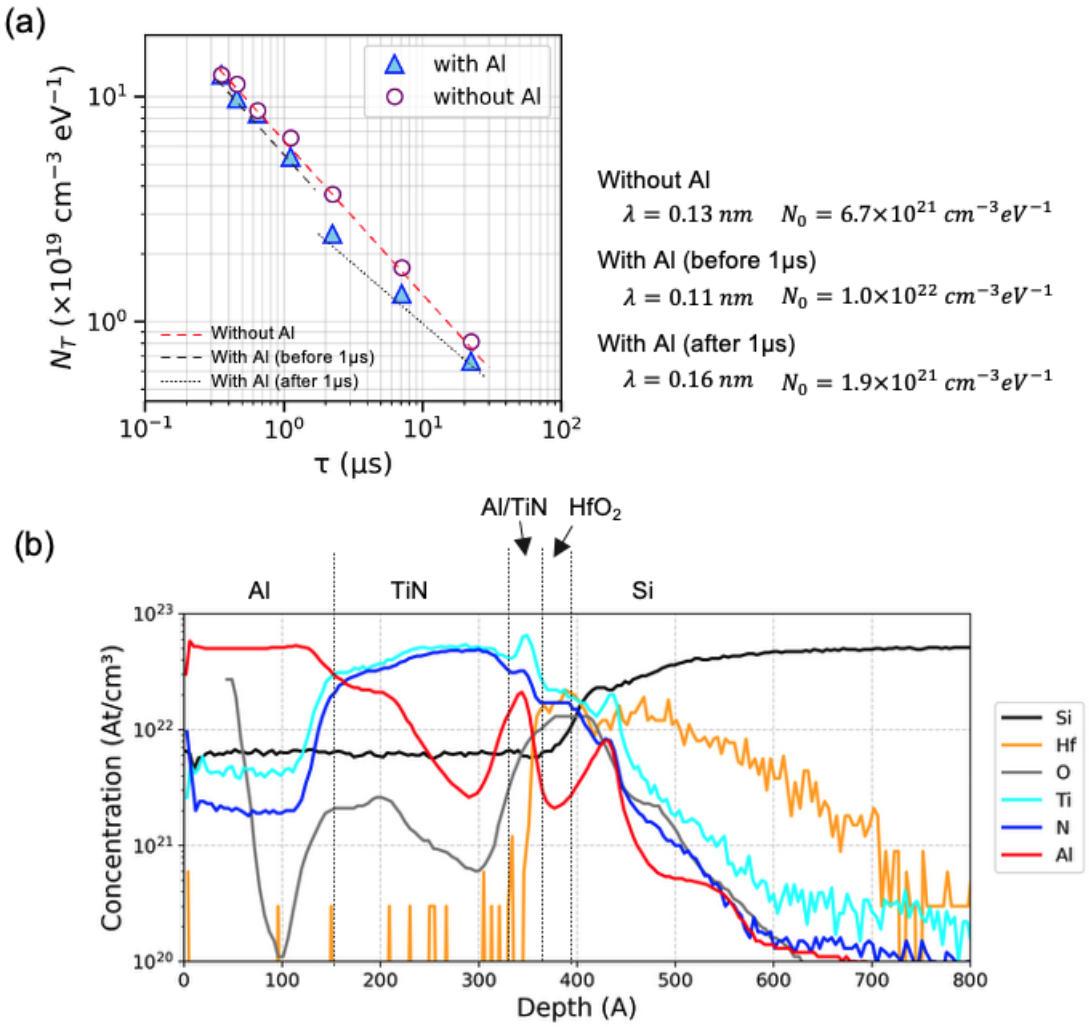


Figure 3.3: (a) Volumetric trap density (N_T) vs. period of C–V measurement frequency for MOSFETs with and without 1 nm ultrathin Al (2.5 nm HfO_2). (b) SIMS depth profiles through the HKMG stack with 1 nm ultrathin Al (2.5 nm HfO_2).

3.4 Impact of Al incorporation on metallic gate/high-k electron energy barrier height and effective gate work function

MOSCAPs with a relatively thick HfO_2 film (8 nm) were used to determine the electron energy barrier height at the metal-oxide interface, from I–V measurements in the Fowler–

Nordheim (F–N) tunneling regime of operation. Figures 3.4 (a) and 3.4 (b) show the measured gate current as a function of gate voltage before and after FGA, respectively, with inset plots showing how the barrier height values were extracted [50].

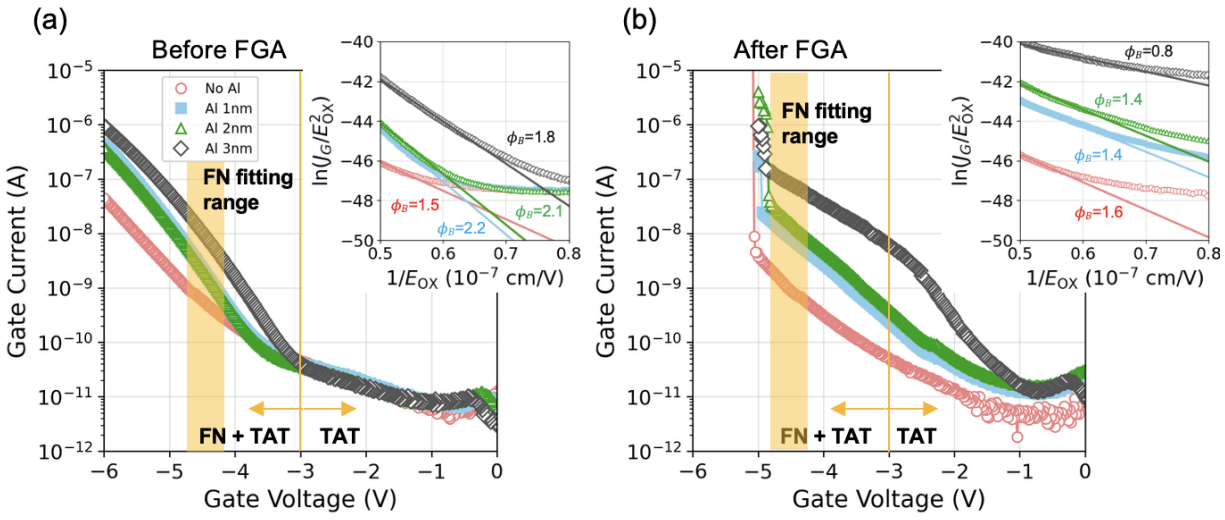


Figure 3.4: Measured gate current vs. gate voltage of a MOSCAP with 8 nm HfO_2 gate dielectric in the FN regime of operation, for different thicknesses of ultrathin Al: (a) before FGA and (b) after FGA. Insets plot $\ln(J_G/E_{OX}^2)$ vs. $1/E_{OX}$, from which the slope is used to extract the effective electron tunneling barrier height. FN fitting was performed over V_G range of -4.3 V to -4.8 V for both before FGA and after FGA.

Figure 3.5 (a) shows how the electron barrier height at the metal/oxide interface decreases with increasing Al layer thickness, as well as the impact of FGA. Before FGA, the barrier height decreases from 2.2 eV to 1.8 eV for ultrathin Al thickness from 1 nm to 3 nm. The Al 0 nm sample is excluded from this trend because the fixed FN fitting range used for comparison lies below its actual FN tunnelling regime, leading to an inaccurate barrier height extraction. After FGA, the barrier height decreases significantly ranging from 1.6 eV to 0.8 eV as the ultrathin Al thickness increases from 0 nm to 3 nm. In contrast, the flat-band voltage (V_{FB}) exhibits only minor shifts following FGA (Figure 3.5 (b)), although it also decreases with ultrathin Al layer thickness. Notably, the pre-FGA barrier height values agree well with those estimated based on measured V_{FB} values (cf. Figure 3.5 (a)). This suggests that the apparent change in barrier height induced by FGA is not primarily due to a change in the effective gate work function.

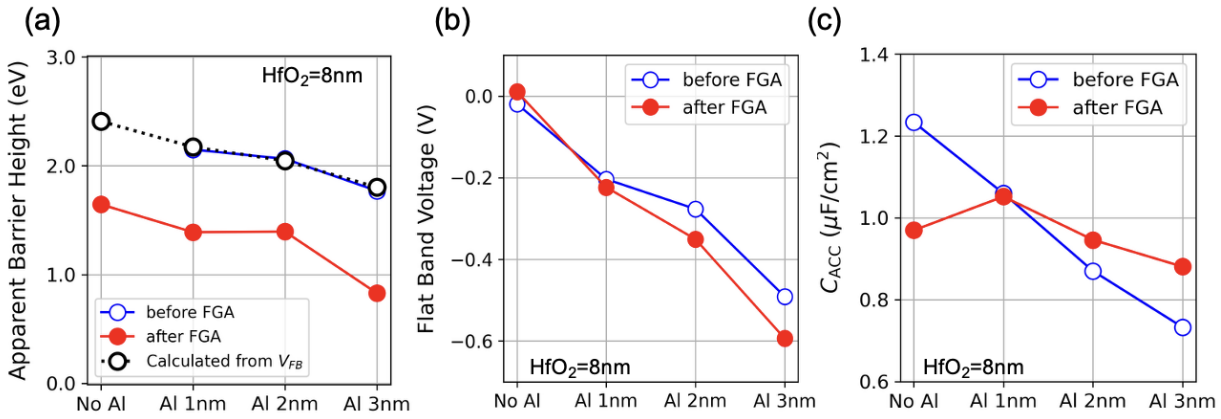


Figure 3.5: (a) Extracted barrier height, (b) flat band voltage (V_{FB}), and (c) accumulation capacitance (C_{ACC}) vs. ultrathin Al thickness for a MOSCAP with 8 nm HfO_2 gate dielectric, before and after FGA.

To elucidate the physical mechanism(s) underlying the observed changes in barrier height and V_{FB}, the MOSCAP samples were analyzed using cross-sectional transmission electron microscopy (XTEM) as shown in Figure 3.6 (e) and (f). Figure 3.6 (a)–(c) shows the elemental depth profiles through the HKMG film stack, obtained via energy-dispersive spectroscopy (EDS). The data reveal that the Al and TiN layers nearest to the HfO_2 interface (positions 3–6 in Figure 3.6 (a)) contain oxygen at concentrations comparable to those in the HfO_2 layer. The electron energy loss spectroscopy (EELS) spectra in Figure 3.7 further confirms that the TiN barrier layer (position 4 in Figure 3.6 (a)) and ultrathin Al layer (position 6 in Figure 3.6 (a)) are in oxidized states—Ti^{4+} [51, 52] and Al^{3+} [53]—in contrast to the metallic TiN (Ti^{3+}) and Al states observed in the upper layers (position 9 in Figure 3.6 (a) for TiN and positions 12, 13 in Figure 3.6 (a) for Al).

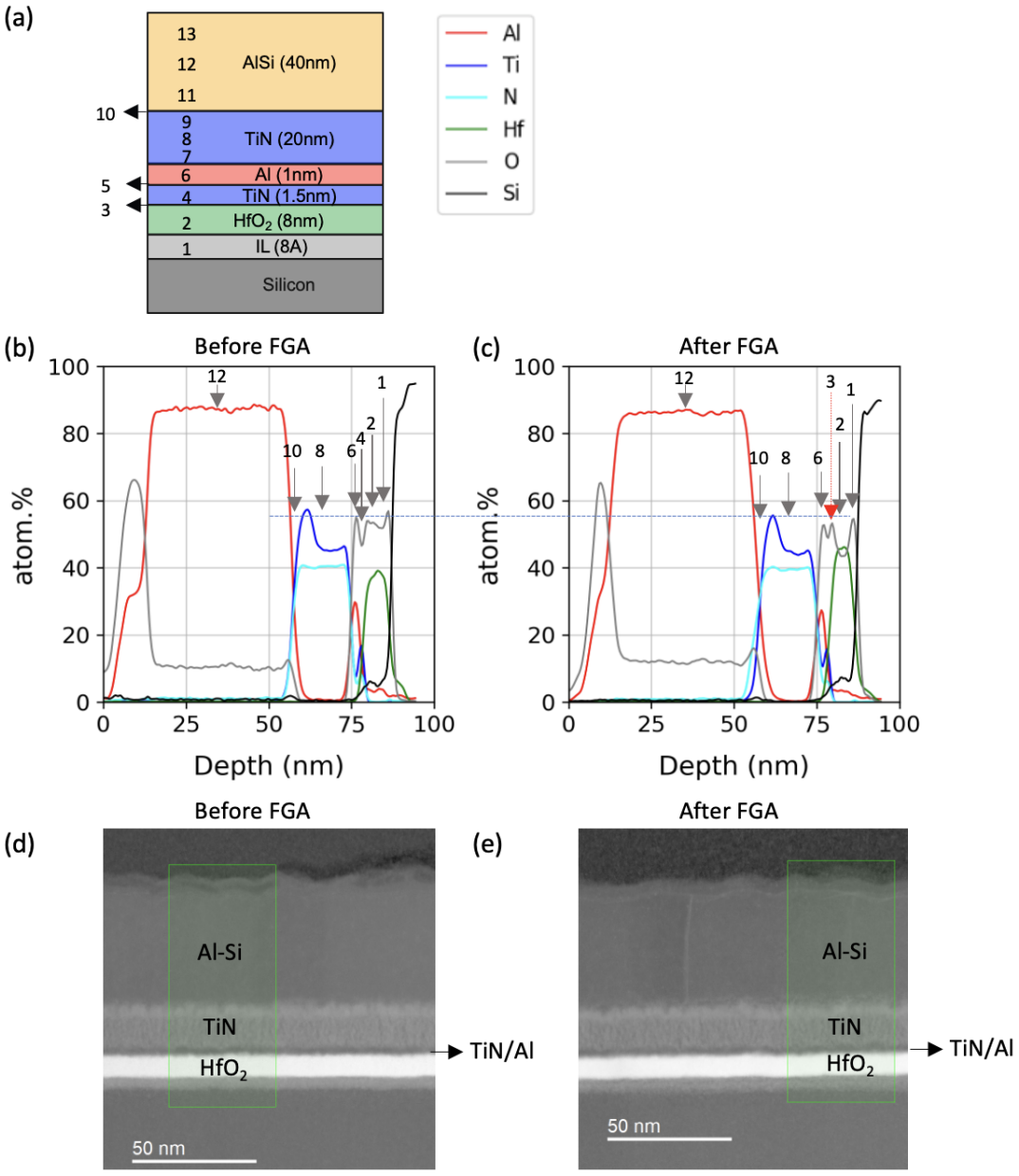


Figure 3.6: (a) MOSCAP HKMG stack with numbered labels for each layer. (b) EDS analysis before FGA. (c) EDS analysis after FGA. (d) XTEM used for EDS analysis before FGA. (e) XTEM used for EDS analysis after FGA.

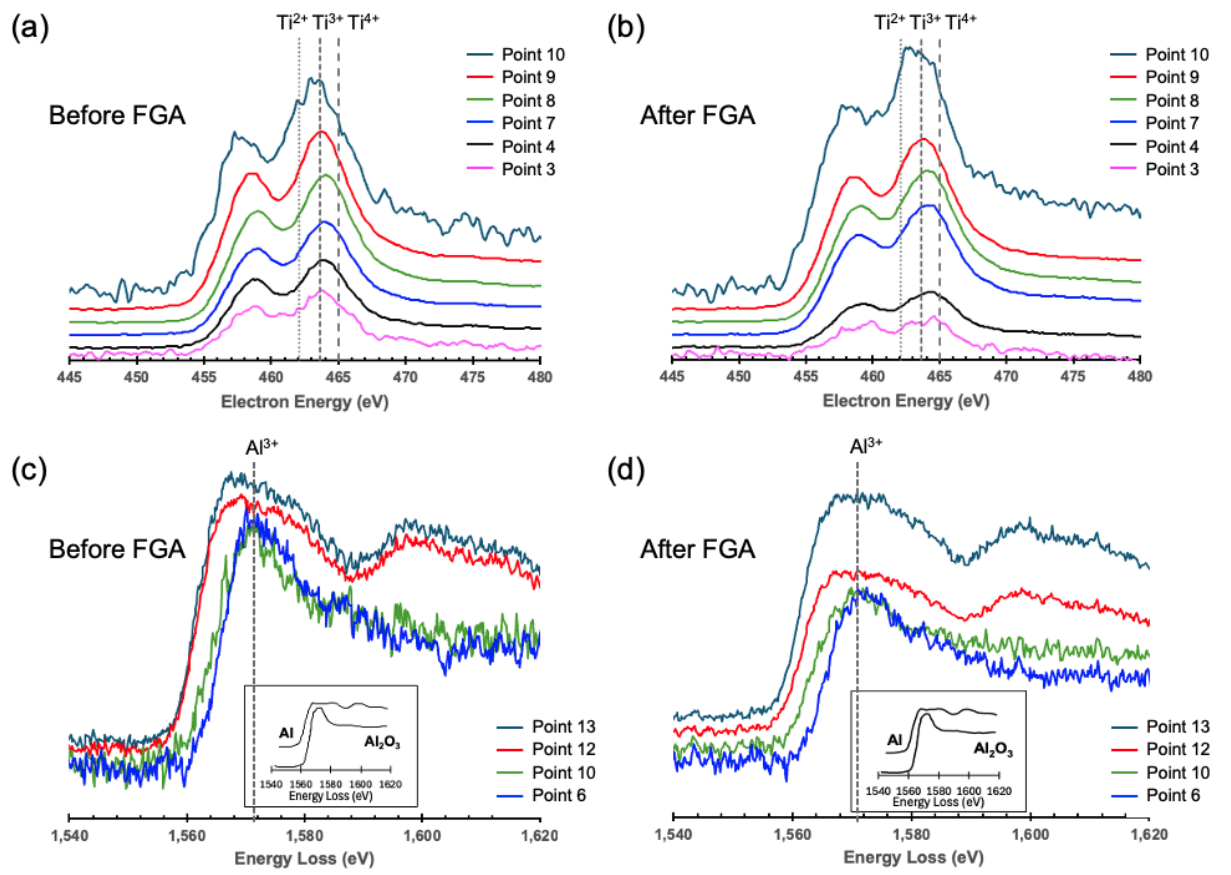


Figure 3.7: MOSCAP (8 nm HfO_2, 1 nm ultrathin Al) chemical analyses by EELS. (a) Ti $L_{2,3}$-edge spectra, before FGA. (b) Ti $L_{2,3}$-edge spectra, after FGA. (c) Al K-edge spectra, before FGA. (d) Al K-edge spectra, after FGA. Reference peak positions for Ti^{2+}, Ti^{3+}, and Ti^{4+} [51, 52], as well as Al and Al_2O_3 spectra [53], are indicated in the plots.

Figure 3.8 illustrates the physical mechanism responsible for the observed changes in barrier height and V_{FB} with the incorporation of Al in the metallic gate. For a previously reported TiN/TiAlC gate film [54], the TiN exists as TiON, which exhibits a higher effective work function than TiN. Al in the TiAlC layer scavenges oxygen from the TiON layer, reducing the oxygen concentration and driving the TiON composition toward TiN, which lowers the effective gate work function. In the current study, both Al and Ti at the HfO_2 interface are observed in oxidized form, as confirmed by the XTEM EDS/EELS analysis. As the ultrathin Al layer thickness is increased, the oxygen content in the TiON layer decreases, resulting in a reduced effective gate work function. Figure 3.5 (c) shows that accumulation capacitance (C_{ACC}) decreases with increasing Al thickness, suggesting that a thicker Al layer

promotes greater AlOx formation by scavenging oxygen from the TiON layer.

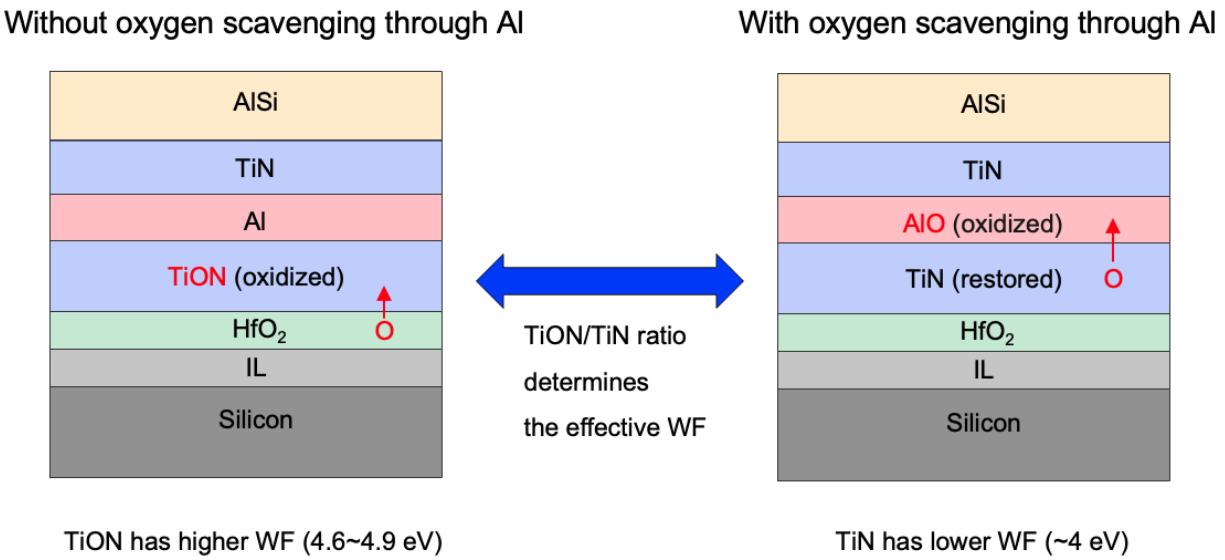


Figure 3.8: Schematic illustration of oxygen scavenging by the ultrathin Al layer within the HKMG stack, reducing the oxygen concentration in TiON layer and thereby shifting its composition toward TiN with a lower effective work function.

It is interesting to note that FGA results in oxygen redistribution as shown by the EDS analyses in Figures 3.6 (b) and 3.6 (c). Specifically, there is less oxygen (i.e. the formation of oxygen vacancies) within the HfO_2 film (position 2) after FGA. Correspondingly, the gate leakage current in the low-field regime increases after FGA for the devices with an ultrathin Al layer (cf. Figure 3.4 (b)), indicative of trap-assisted tunneling (TAT). The reduced apparent barrier height after FGA (Figure 3.5 (a)) is attributed to the increased contribution of TAT, which causes the leakage current to deviate from ideal FN tunnelling and consequently leads to an underestimation of the FN barrier height. In addition, all samples after FGA exhibit dielectric breakdown around -5 V. These results suggest that a reduction in effective gate work function achieved by incorporating Al in the metallic gate comes at a tradeoff of degraded HfO_2 quality, due to oxygen redistribution during FGA.

With oxygen migration from the HfO_2 layer towards the metallic gate, a new oxygen peak at the gate-HfO_2 interface (position 3 in Figure 3.6 (c)) is observed after FGA. At the same time, the oxygen concentration in the ultrathin Al layer (position 6 in Figure 3.6 (c)) is reduced while it is increased in the TiN barrier layer after FGA. This redistribution of oxygen results in slightly higher accumulation capacitance (C_{ACC}) after FGA, as shown in Figure 3.5 (c), for ultrathin Al layer thickness greater than 1 nm. (The 1 nm-thick ultrathin

Al layer is already completely oxidized before FGA due to air exposure, as explained above.) This effect is more pronounced for a thicker ultrathin Al layer. On the other hand, the MOSCAP without an ultrathin Al layer exhibits a reduction in C_{ACC} after FGA, likely due to oxygen transport from HfO_2 into the (TiN) gate (cf. Figure 3.4 (b) showing slightly higher TAT for 0 nm ultrathin Al, after FGA).

3.5 HfO_2 dielectric degradation due to excess Al in the gate film stack

Finally, we investigated the degradation of HfO_2 in MOS devices with a single thick Al layer in the gate film. MOSCAPs with an 8.0 nm-thick HfO_2 layer were fabricated with a 40 nm-thick Al-Si layer (2 % Si) deposited via DC sputtering on top of a TiN diffusion barrier layer (0 to 6 nm thick) deposited by ALD. Figure 3.9 shows measured C–V and gate leakage characteristics before and after FGA for six devices on each wafer. As can be seen from Figure 3.9 (a), devices without a TiN barrier layer fail to reach the expected accumulation capacitance under high negative gate bias, indicating a high concentration of charge traps in the gate dielectric. Correspondingly, high gate leakage is observed in Figure 3.9 (c) and Figure 3.9 (d). Both the accumulation capacitance degradation and the number of devices exhibiting low gate leakage decrease with increasing TiN barrier layer thickness.

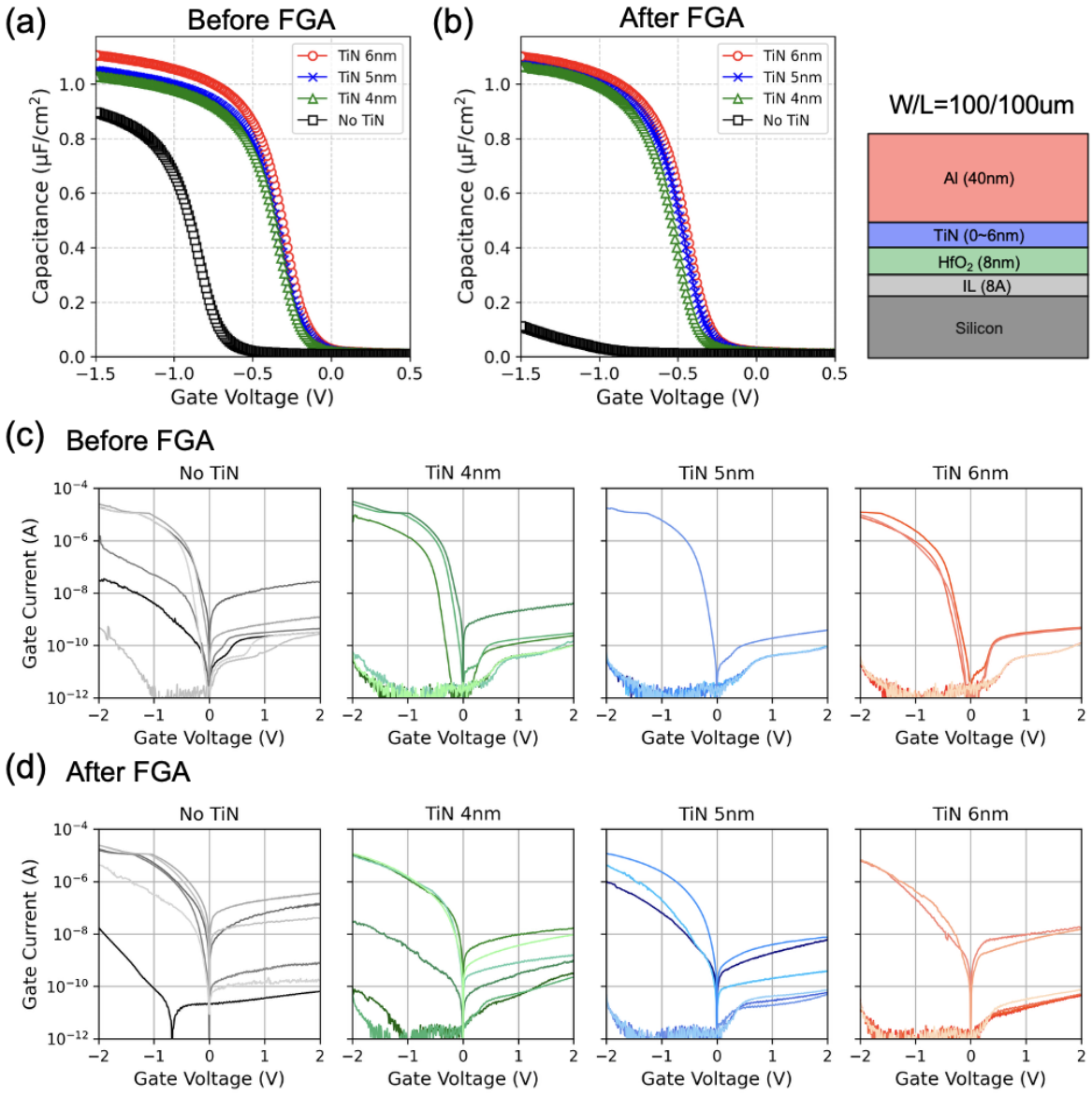


Figure 3.9: Measured C–V curves for MOSCAPs with 40 nm Al gate, 8 nm HfO_2, and different interfacial TiN barrier layer thicknesses: (a) before FGA and (b) after FGA. Measured gate current vs. gate voltage for six MOSCAPs each: (c) before FGA and (d) after FGA.

Figure 3.10 compares elemental depth profiles obtained via secondary ion mass spectrometry (SIMS) for MOSCAP samples before and after FGA, with and without a 6.0 nm-thick TiN diffusion barrier. Even before FGA (Figure 3.10 (a)), upward diffusion of Hf and Si into the metal gate is observed in the absence of a TiN barrier, and interdiffusion of Hf and Si

is also evident at the HfO_2/Si interface. FGA further enhances interdiffusion (Figure 3.10 (b)). The TiN layer effectively limits the diffusion of Hf and Si to the TiN/HfO_2 interface (Figure 3.10 (c) and Figure 3.10 (d)). The TiN barrier also suppresses interdiffusion at the HfO_2/Si interface. These results demonstrate that a TiN diffusion barrier is essential when integrating Al into the metallic gate film, to maintain dielectric integrity while providing for a means to tune the effective gate work function. The increased TAT leakage seen after FGA (cf. Figure 3.4 (b)) is likely due to insufficient TiN thickness to prevent intermixing between Al and HfO_2.

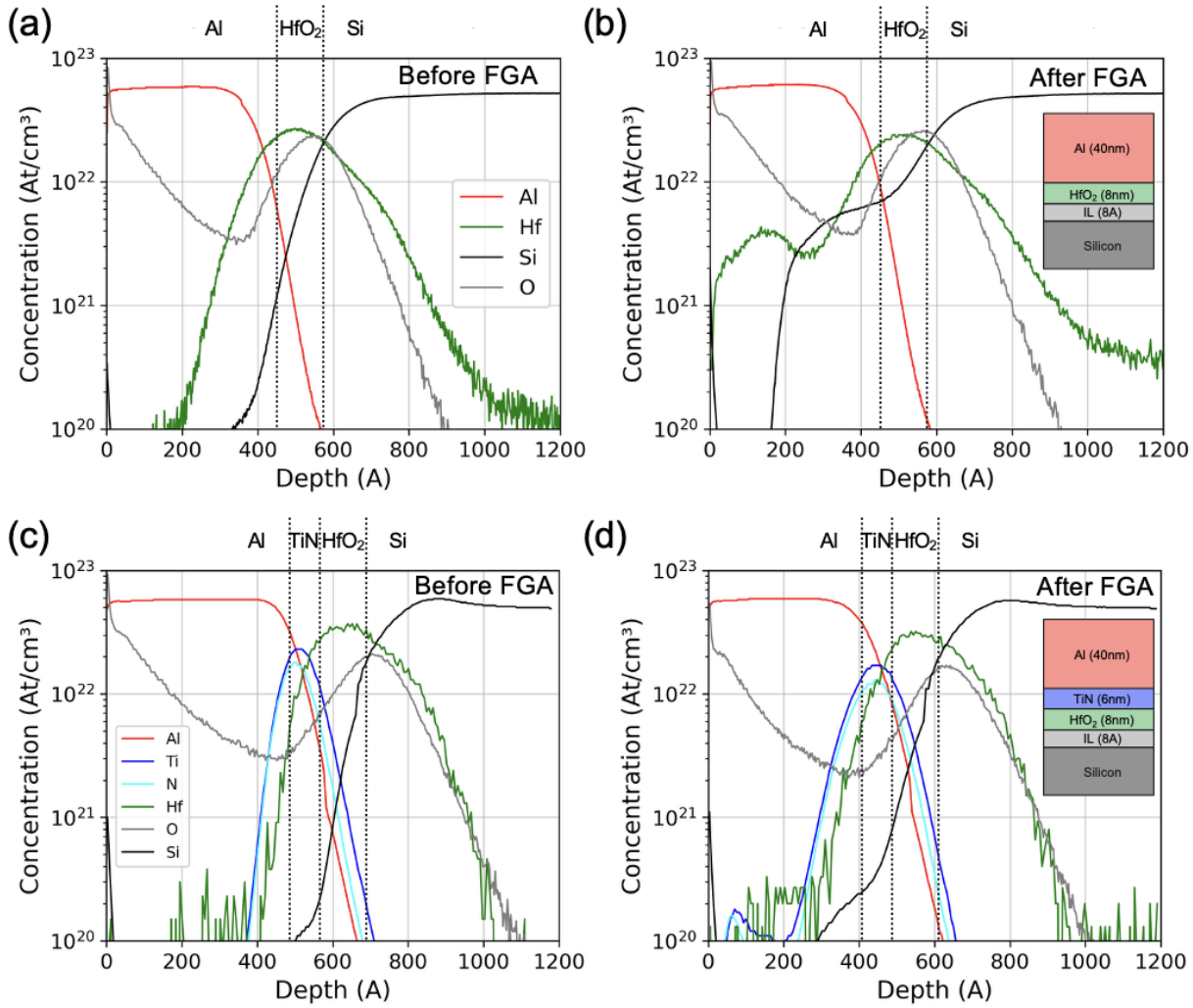


Figure 3.10: SIMS depth profiles through the HKMG stack with 40 nm Al gate and 8nm HfO2, without vs. with an interfacial 8nm TiN barrier layer: (a) without TiN before FGA and (b) after FGA; (c) with TiN before FGA and (d) after FGA.

3.6 Conclusion

The effective gate work function of HfO_2-based MOS devices can be tuned by incorporating an ultrathin aluminum layer within the metallic gate film, causing the flat-band voltage to decrease with increasing Al layer thickness. This effect is attributed to oxygen uptake by the Al layer, which lowers the oxygen content in the barrier TiON layer, shifting its composition toward TiN. However, the incorporation of Al within the metallic gate film can

also result in Al within the HfO_2, which introduces an additional electron trap state near the oxide interfaces. Furthermore, FGA promotes oxygen redistribution from the HfO_2 film to the metallic gate film, leaving behind oxygen vacancies in the HfO_2 and resulting in TAT leakage and degraded dielectric breakdown voltage. A sufficiently thick TiN diffusion barrier is necessary to prevent upward diffusion of Hf and Si (toward the Al gate film), thereby preserving dielectric integrity.

Table 3.1: Summary of key observations and physical mechanisms.

Topic	Key Observation	Physical Mechanism
1. Al induced effective WF modulation	V_{FB} decreases with increasing Al thickness	Al scavenges oxygen from TiON, shifting the TiON/TiN ratio toward TiN with a low WF
2. Al induced trap generation	Devices with Al exhibit higher D_{IT} with larger tunneling attenuation length	Al incorporation into HfO_2 introduces additional electron trap states near the oxide/metal interface
3. FGA induced oxygen redistribution	Low field leakage increases after FGA	Oxygen migrates from HfO_2 into the metallic gate, leaving oxygen vacancies in HfO_2
4. Role of TiN diffusion barrier	Devices without sufficient TiN show severe leakage and degraded C–V characteristics	TiN suppresses upward diffusion of Hf and Si into the Al gate

Chapter 4

Work Function Engineering with OI Technology for advanced CMOS devices

4.1 Introduction

An oxygen inserted (OI) layer is a partial oxygen monolayer formed in crystalline silicon during epitaxial growth, using Mears Silicon Technology (MST) [55, 56]. The details of this growth process are not described in this dissertation because they are proprietary. Figure 4.1 (a) illustrates the atomic configuration of the OI layer in silicon. The inserted oxygen atoms occupy interstitial sites and form covalent bonds with neighboring silicon atoms. Figure 4.1 (b) presents the O 1s X-ray photoelectron spectroscopy (XPS) spectrum of the OI layer [57]. The spectrum indicates that the oxygen atoms are in an intermediate chemical state between the covalent oxygen in SiO_2 and an approximately 0.5 eV ionized oxygen state.

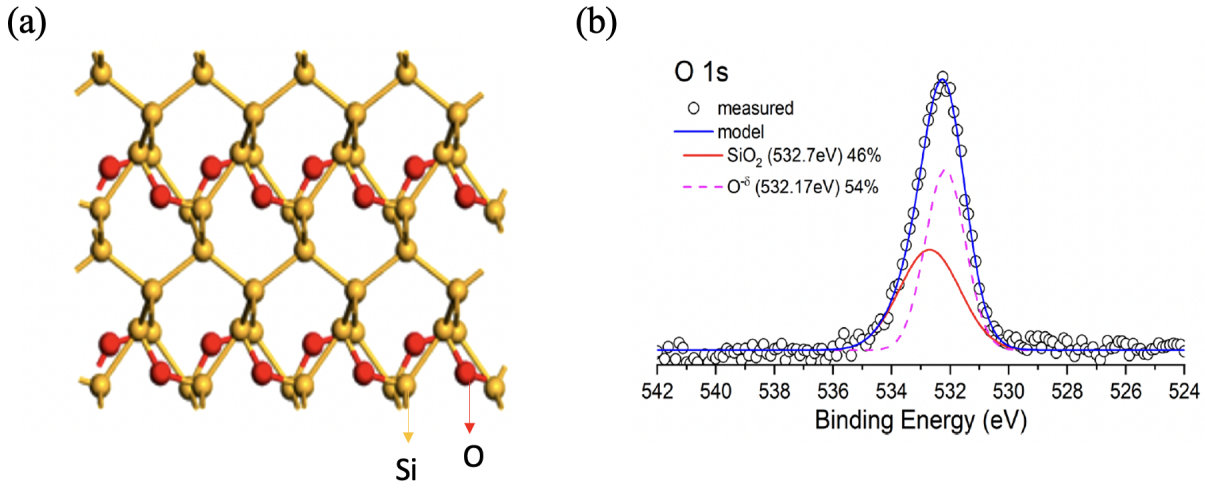


Figure 4.1: (a) Schematic atomic configuration of OI layer in silicon. (b) O 1s XPS spectrum of OI layer in silicon. Adapted from [57].

Previous studies have reported that the incorporation of an OI layer in the body of bulk-silicon n-channel MOSFETs with poly-Si/SiO_2 gate stack improves electron field-effect mobility [58]. As shown in Figure 4.2, the effective electron mobility increases by approximately 10 % at 300 K and 50 % at 4 K compared with the control device. The enhancement is most pronounced at high effective electric fields, suggesting that the OI layer results in reduced surface roughness scattering.

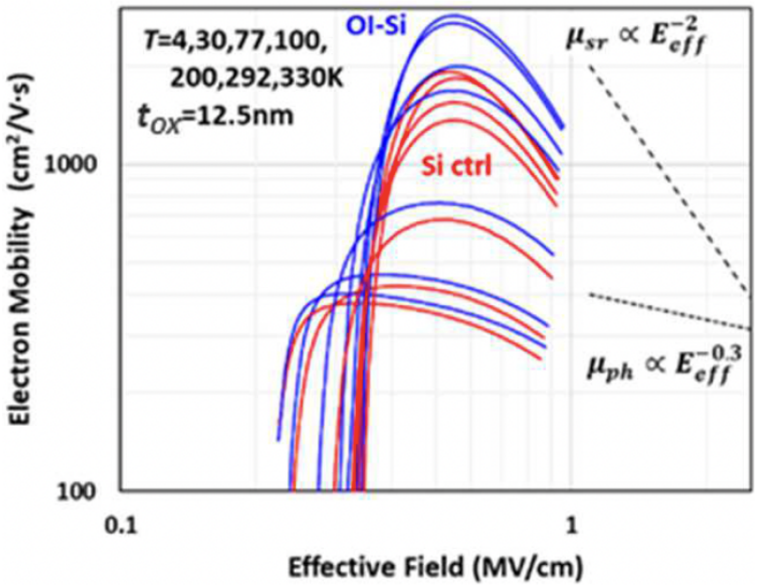


Figure 4.2: Electron mobility as a function of effective transverse electric field for control and OI bulk-Si NMOSFETs at various temperatures. Adapted from [58].

To investigate the mechanism responsible for the apparent surface roughness reduction, an isotope tracing experiment was conducted [58]. Four OI layers were formed on a bulk-silicon wafer using the ^{18}O isotope, followed by dry oxidation at 900 °C in a ^{16}O ambient. Because the SiO_2 layer was grown using only ^{16}O, the OI layer and the thermally grown oxide could be distinguished by their oxygen isotopes. As shown in Figure 4.3, secondary ion mass spectrometry (SIMS) analysis revealed that ^{18}O was detected within the SiO_2 layer, while ^{16}O was observed at the original OI layer position after oxidation. These results indicate that oxygen exchange occurs between the OI layer and the growing SiO_2 during thermal oxidation, which is believed to contribute to the reduction of surface roughness in silicon containing the OI layers.

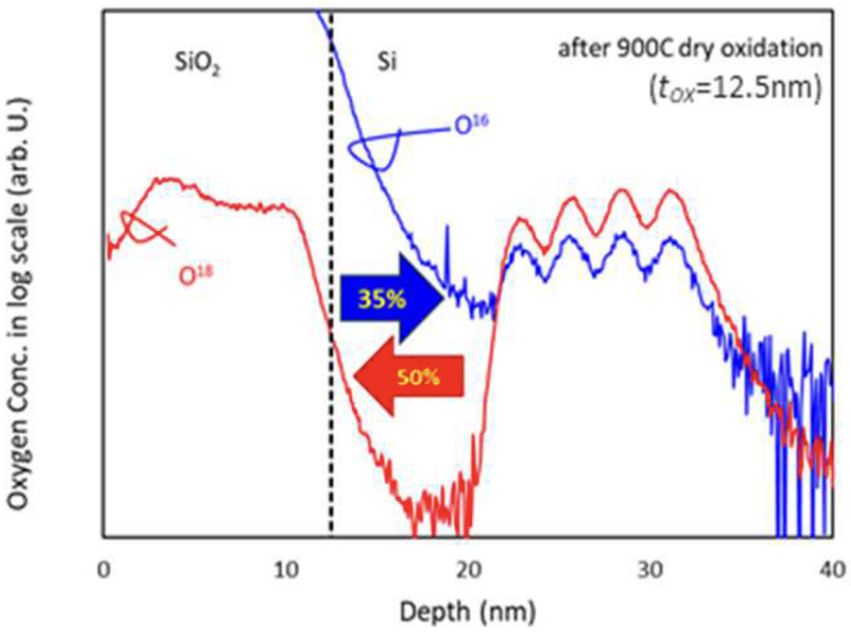


Figure 4.3: Depth profiles of ^{16}O and ^{18}O after dry oxidation at 900 °C. Four OI layers were formed using ^{18}O, followed by dry oxidation in a ^{16}O ambient to grow the SiO_2 layer. Adapted from [58].

Moreover, a recent study from San José State University analyzed the Si/SiO_2 interface using high-resolution TEM images and extracted the interface roughness power spectral density (PSD) [59]. The results showed that OI silicon exhibits lower interface roughness at low spatial frequencies, which was identified as the origin of the enhanced surface-roughness-limited electron mobility.

In addition to reducing surface roughness scattering, the OI layer has also been reported to suppress Coulombic scattering in HKMG MOSFETs as described in Figure 4.4 [60]. During thermal processing of HKMG devices, oxygen is transported from the HfO_2 dielectric to the Si substrate, leading to the formation of an interfacial SiO_2 layer. This oxygen transport generates oxygen vacancies in the HfO_2 layer, which become positively charged after electron transfer. The positively charged oxygen vacancies and negatively charged oxygen ions across the interfacial SiO_2 layer form dipoles that act as Coulomb scattering centers for channel carriers. When an OI layer is incorporated into the Si substrate, oxygen required for interfacial SiO_2 formation can be supplied directly from the OI layer instead of the HfO_2 dielectric. Consequently, fewer oxygen vacancies are generated in HfO_2, reducing

the density of interfacial dipoles and thereby suppressing Coulomb scattering.

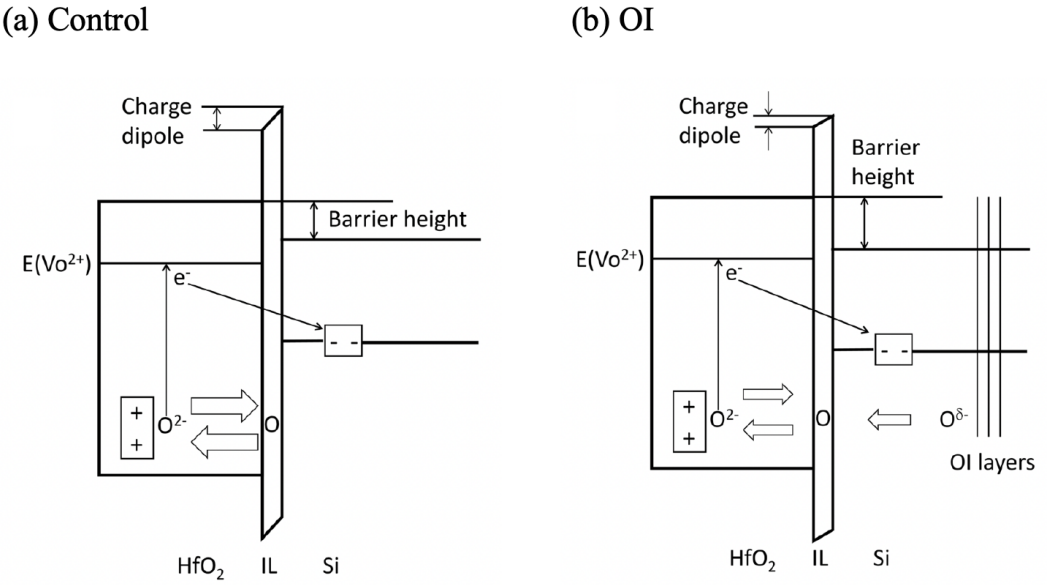


Figure 4.4: Band diagrams of HfO_2, SiO_2 and Si illustrating charge-dipole formation in (a) the control device and (b) the OI device. Adapted from [60].

Beyond mobility enhancement, MOSFETs fabricated on OI silicon have also been reported to exhibit a lower flat-band voltage than control devices. However, the physical origin of this voltage shift remains unclear. In this study, electrical characterization and chemical analysis are performed to elucidate its underlying mechanism.

4.2 Experimental

Three types of MOS capacitor (MOSCAP) samples were fabricated, as illustrated in Figure 4.5. The first sample, referred to as the Control, was fabricated on a standard p-type Si substrate. A chemical oxide interfacial layer was first formed by immersion in heated H_2O:NH_4OH:H_2O_2 (200:1:1 by volume). Subsequently, HfO_2 was deposited by atomic layer deposition (ALD) at 250 °C using tetrakis(dimethylamido)hafnium(IV) (TDMAHf, $Hf(NMe_2)_4$) as the Hf precursor and H_2O as the oxidant. Before the HfO_2 deposition, five H_2O pulse cycles were applied to improve the quality of the chemical oxide. An 8.0 nm-thick HfO_2 film was deposited using 90 ALD cycles. A 3.5 nm TiN barrier layer was then deposited by plasma-enhanced ALD at 300 °C using tetrakis(dimethylamido)titanium

(TDMAT, $Ti(NMe_2)_4$) as the Ti precursor and H_2/N_2 plasma as the reactant. The TiN deposition consisted of 30 ALD cycles. Finally, a 40 nm tungsten (W) gate electrode was deposited using a 5 kHz pulsed-DC sputtering system.

The second sample, referred to as the OI sample, was fabricated on an OI Si substrate. Five sequential OI layers were embedded in the Si substrate, with the uppermost OI layer located 10 nm below the Si surface. The chemical oxide interfacial layer, HfO_2 gate dielectric, TiN barrier layer, and W gate electrode layers were formed using the same process flow as that for the Control sample.

The third sample, referred to as the Al sample, was fabricated on a standard Si substrate. Unlike the control sample, the gate stack included an ultrathin Al layer inserted between two TiN layers. The bottom TiN layer was deposited by ALD using 12 cycles, resulting in a thickness of 1.5 nm. Subsequently, a 2.0 nm Al layer was deposited by electron-beam evaporation, followed by deposition of a 10 nm top TiN layer and a 40 nm W gate electrode by sputtering.

The gate stacks were patterned by photolithography and reactive ion etching (RIE), with the etch process stopped on the HfO_2 gate dielectric. The fabricated MOSCAPs had an area of 100 μm by 100 μm.

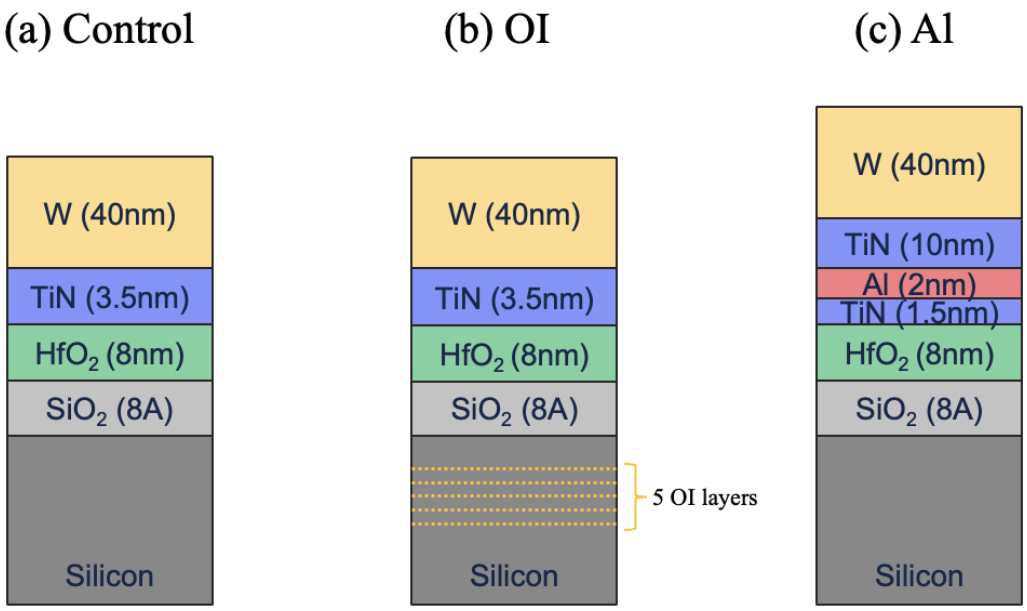


Figure 4.5: Schematic illustrations of the three MOSCAP structures: (a) Control with a standard Si substrate, (b) OI with an OI Si substrate, and (c) Al with an Al layer incorporated into the HKMG gate stack.

4.3 Results & Analysis

Figure 4.6 shows the C–V characteristics of the MOSCAPs measured using 100 kHz small-signal frequence and using bidirectional bias voltage sweeps. The extracted V_{FB} values are 0.37 V for the Control sample, -0.24 V for the OI sample, and -0.39 V for Al sample, demonstrating that both OI and the ultrathin Al layer effectively shift the V_{FB} toward more negative values.

Although the Al sample exhibits a slightly lower V_{FB} than the OI sample, the OI sample shows a steeper transition between accumulation to depletion. In addition, the hysteresis voltage determined from the difference between the forward and reverse voltage sweeps is smaller for the OI sample than for the Control and Al samples. These observations suggest that the OI sample has a lower interface trap density [61], indicating that OI provides an additional benefit beyond V_{FB} engineering.

A slight difference in the accumulation capacitance (C_{ACC}) is also observed among the three samples, despite the dielectric layers being fabricated with identical processes. Before forming gas annealing (FGA), all three samples exhibited the same accumulation capaci-

tance of approximately 1.32 μF/cm^2. The difference observed after FGA therefore suggests that the annealing process slightly modified either the effective dielectric thickness or the effective dielectric constant of the gate dielectric stack, likely as a result of different degrees of intermixing or oxidation.

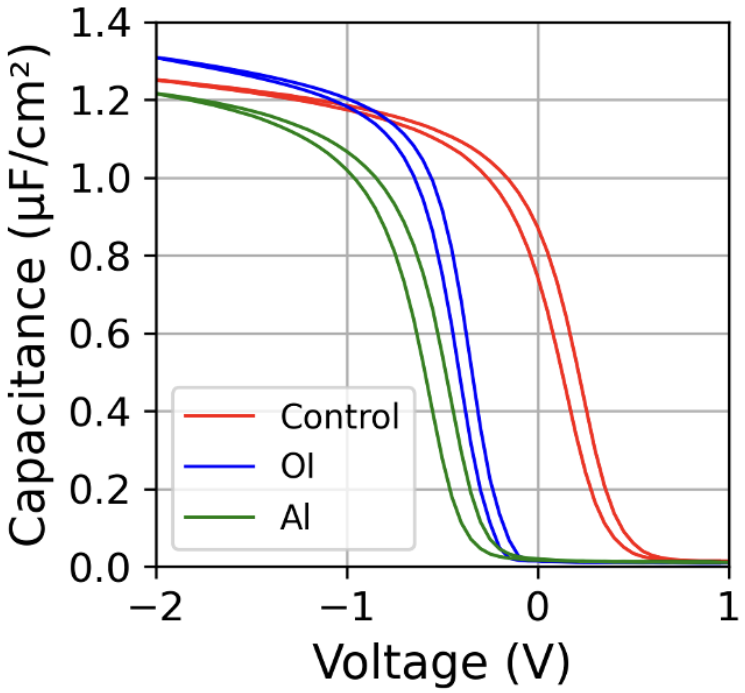


Figure 4.6: C–V characteristics of the Control, OI, and Al MOSCAPs measured with a small-signal frequency of 100 kHz using bidirectional bias voltage sweeps after forming gas annealing (FGA).

To investigate the origin of the negative V_{FB} shift induced by OI, secondary ion mass spectrometry (SIMS) was used to compare the elemental depth profiles of the HKMG stacks in the Control and OI samples, as shown in Figure 4.7. Compared with the Control sample, the OI sample exhibits more confined elemental profiles with more abrupt interfaces between adjacent layers. These observations indicate that OI layer suppresses elemental interdiffusion within the HKMG stack.

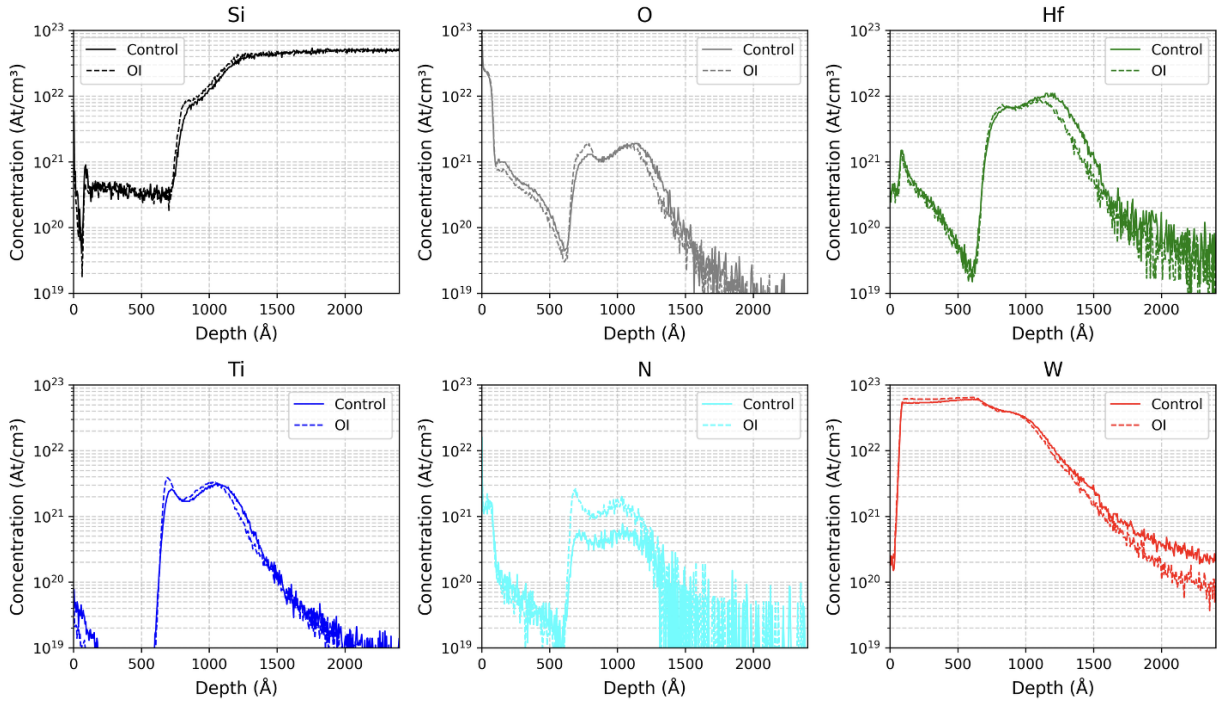


Figure 4.7: SIMS depth profiles of Si, O, Hf, Ti, N, and W for the Control and OI samples.

Figure 4.8 presents the same SIMS data as Figure 4.7, with all elemental depth profiles overlaid for the Control and OI samples to facilitate comparison within each HKMG stack. In the Control sample, the oxygen profile extends broadly across both the HfO_2 and TiN layers. In contrast, the oxygen profile in the OI sample is much more localized at the HfO_2/TiN interface. As discussed in Chapter 3, oxygen diffusion from the HfO_2 layer into TiN promotes TiON formation, resulting in a positive shift in V_{FB}. The broad oxygen distribution observed in the Control sample is consistent with oxygen diffusion into the TiN layer and the associated TiON formation. In contrast, the confined oxygen distribution in the OI sample suggests that oxygen diffusion into TiN is effectively suppressed because of reduced HKMG intermixing. Consequently, TiN oxidation is mitigated, resulting in the observed negative shift in V_{FB}.

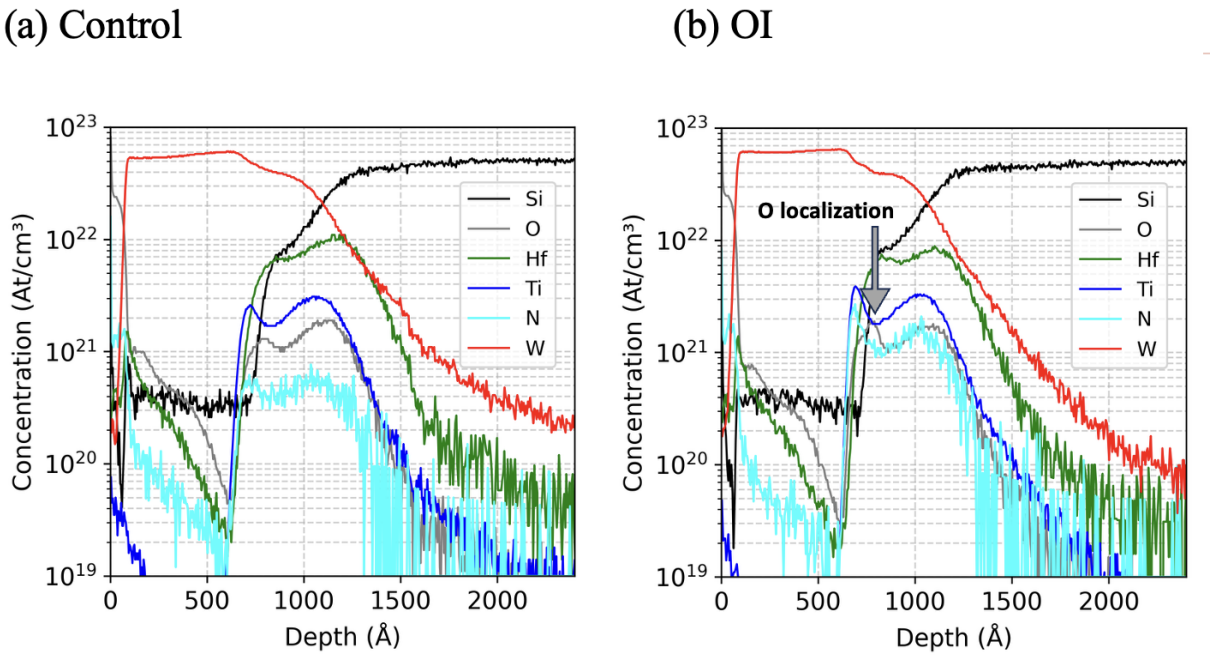


Figure 4.8: Overlaid SIMS depth profiles of Si, O, Hf, Ti, N, and W for the (a) Control and (b) OI samples, highlighting differences in oxygen distribution within the HKMG stacks.

To further investigate the oxidation state of TiN in the Control and OI samples, oxygen K-edge electron energy loss spectroscopy (EELS) was performed at six different positions throughout the HKMG stack. Figure 4.9 compares the normalized oxygen K-edge spectra obtained from the Control and OI samples at each location. At both the top and bottom regions of TiN layer, the Control sample exhibits a stronger oxygen-related signal than the OI sample, indicating a greater extent of oxygen related bonding within the TiN layer [62]. In contrast, the oxygen K-edge spectra measured at the W/TiN interface, TiN/HfO_2 interface, HfO_2 layer and HfO_2/IL interface show little difference between the two samples, suggesting that the oxygen state in these regions remains nearly identical.

The SIMS results in Figure 4.7 and 4.8 demonstrate that oxygen diffusion into the TiN layer is reduced in the OI sample. Consistent with this observation, the EELS analysis reveals a weaker oxygen-related signal within the TiN layer, indicating a lower degree of TiN oxidation. Together, these results suggest that the OI layer suppresses TiON formation by reducing HKMG intermixing and limiting oxygen diffusion into the TiN layer.

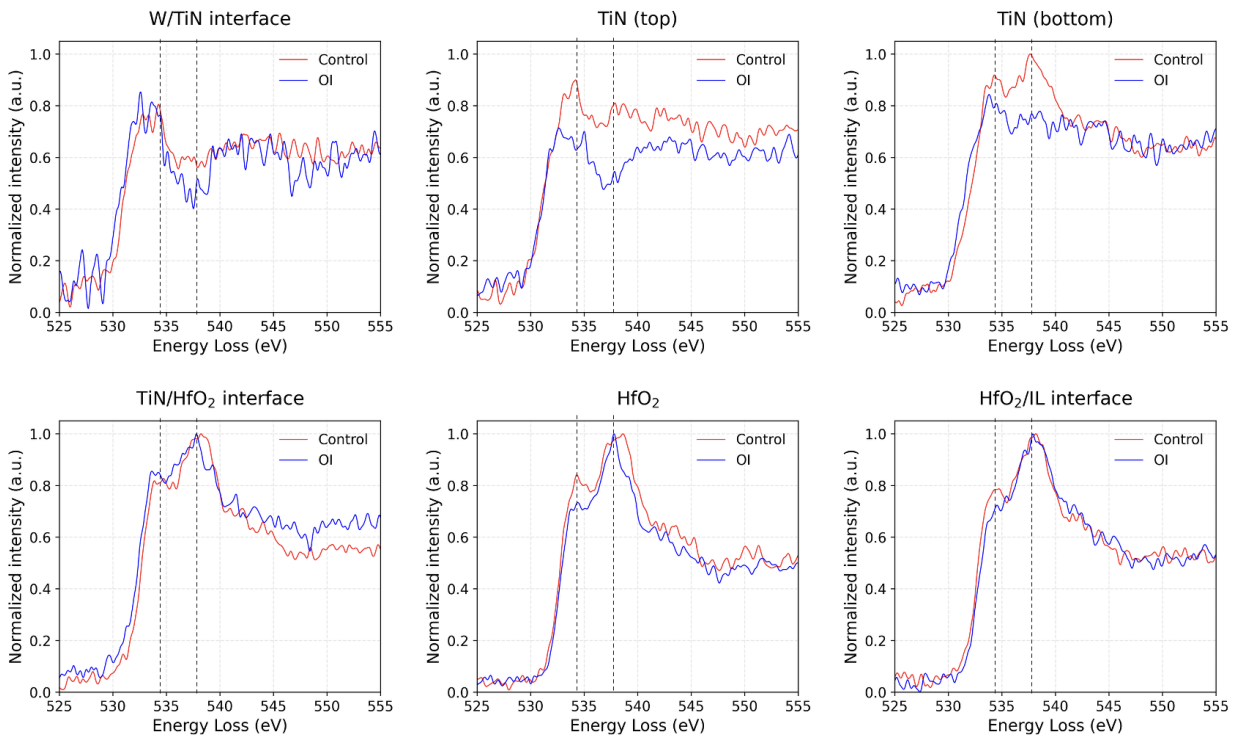


Figure 4.9: Comparison of normalized oxygen K-edge EELS spectra measured at six locations within the HKMG stack for the Control and OI samples.

Lastly, Figure 4.10 compares the Si $L_{2,3}$-edge EELS spectra of the Control and Al samples at three different locations near the gate dielectric. At the HfO_2/SiO_2 interface, the Control sample exhibits a more pronounced Si^{4+} feature than the Al sample, indicating a higher degree of oxide stoichiometry at the interface [63]. In addition, the Si^0 feature at the SiO_2/Si interface and within the Si substrate is more pronounced in the Control sample [63]. These observations suggest that oxygen scavenging by the ultrathin Al layer modifies the stoichiometry of the gate dielectric and influences the local silicon bonding environment near the Si substrate.

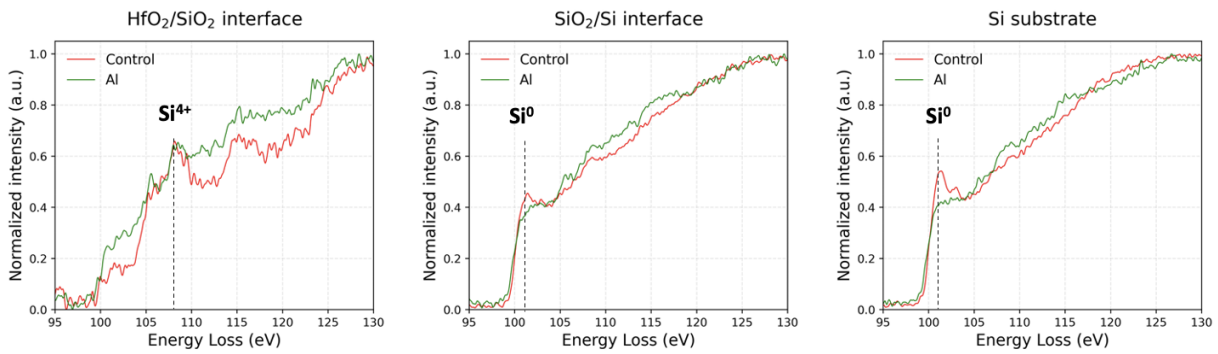


Figure 4.10: Comparison of normalized Si $L_{2,3}$-edge EELS spectra acquired at the HfO_2/SiO_2 interface, SiO_2/Si interface, and Si substrate for the Control and Al samples.

4.4 Conclusion

In summary, OI incorporation in the Si substrate provides for a significant reduction in effective V_{FB}, comparable to that achieved by Al incorporation within the metallic gate. However, the underlying mechanisms are fundamentally different. The OI sample suppresses HKMG intermixing, thereby reducing TiON formation. In contrast, the ultrathin Al layer lowers V_{FB} by scavenging oxygen from the TiON layer and converting it back to TiN.

Compared to ultrathin Al incorporation, the OI substrate approach offers additional advantages. First, it has been reported to improve effective carrier mobility. Second, the OI substrate approach provides for lower interface and dielectric trap densities, as reflected in the smaller C–V hysteresis and steeper accumulation-to-depletion transition observed in this work.

Despite these advantages, several challenges remain before OI can be adopted as a practical V_{FB} engineering technique. Most importantly, the primary process parameter responsible for V_{FB} modulation must be identified and experimentally validated to precisely tune V_{FB}. In addition, the OI approach has so far has been demonstrated only in bulk silicon devices, and its effectiveness in advanced GAAFET technologies with silicon channels as thin as approximately 5 nm has yet to be experimentally validated. Therefore, future studies should evaluate the effectiveness of the OI approach as a function of channel thickness in GAAFET structures.

Chapter 5

SiGe/Si Heterojunction Drain Transistor for Faster 3D NAND Flash Memory Erase

5.1 Introduction

The primary strategy for increasing the storage capacity of a 3D NAND flash memory array is to stack more gates (cells) vertically (in the z-direction) [64]. Figure 5.1 shows the cross-sectional view of a single 3D NAND flash string. In practice, the string is formed by conformally depositing in sequence the oxide-nitride-oxide (ONO) layers of blocking oxide (Box), charge trap nitride (CTN) and tunnel oxide (Tox), followed by the poly-Si channel and inner oxide layers, after a vertical hole is cut through the gate electrode/insulating dielectric multi-layer stack. The plates (shown in purple) surrounding the string correspond to metal gates, and the spaces between the metal gates are filled with dielectric layers. Currently in industry, more than 400 gates can be stacked within a single string.

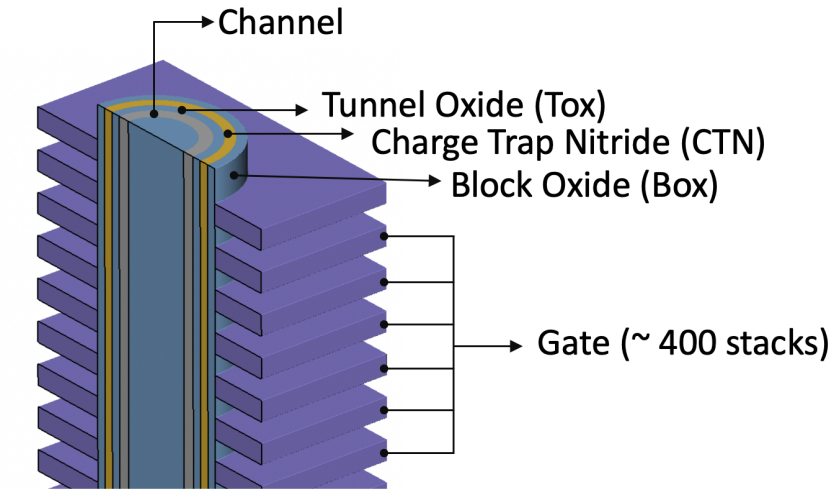


Figure 5.1: Schematic cross-section of the 3D NAND string structure.

There are two main operation modes in NAND Flash: program and erase. Program operation stores electrons in the charge trap nitride layer by applying a positive voltage to the gate while grounding the channel. This creates a high electric field between the channel and the gate, allowing electrons in the channel to move to the charge trap nitride through Fowler–Nordheim (F-N) tunneling, as shown in Figure 5.2 (a). The stored electrons increase the threshold voltage (V_T) of the cell, corresponding to a logic '0'. In contrast, the erase operation removes the electrons from the charge trap nitride layer and returns them to the channel, by applying positive voltage to the channel while keeping the gate at 0 V, as illustrated in Figure 5.2 (b). After the erase operation, the V_T decreases, corresponding to a logic '1'.

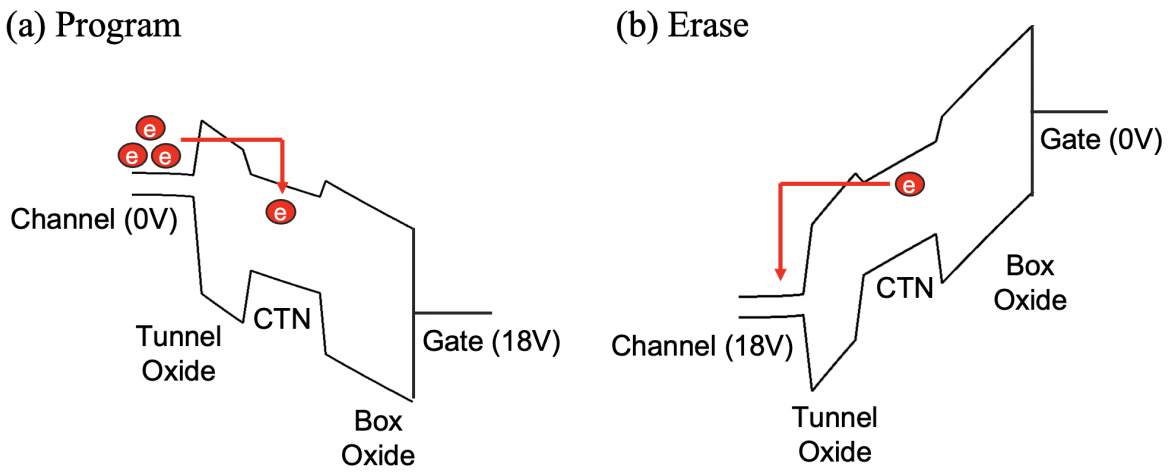


Figure 5.2: Energy band diagrams illustrating (a) the program operation and (b) the erase operation in 3D NAND flash memory.

There are two major erase schemes in NAND flash memory: bulk erase and gate-induced drain leakage (GIDL) erase. The main difference lies in the method of applying a positive voltage to the channel. Bulk erase is a straightforward method in which a positive voltage is directly applied to the channel. However, this approach requires a p-type well (or p-well) beneath the NAND string to transfer the positive voltage to the channel.

On the other hand, GIDL erase utilizes holes generated by band-to-band tunneling (BTBT) at the drain junction. Figure 5.3 illustrates the hole generation mechanism in GIDL erase. Figure 5.3 (a) shows a simplified 3D NAND string consisting of three drain transistors at the top of the string, three cell transistors in the center, and three source transistors at the bottom. Figure 5.3 (b) illustrates the energy band diagram along the channel near the bit line (BL) and drain transistors during GIDL operation. When a positive bias is applied to the n-type bit line (BL) and a negative bias is applied to the gate of the drain transistors, a strong reverse bias is established at the n-type region and the undoped channel. This condition induces band-to-band tunneling, generating electron–hole pairs. The generated electrons are collected by the BL, while the holes remain in the channel, thereby boosting the channel potential toward the applied BL voltage.

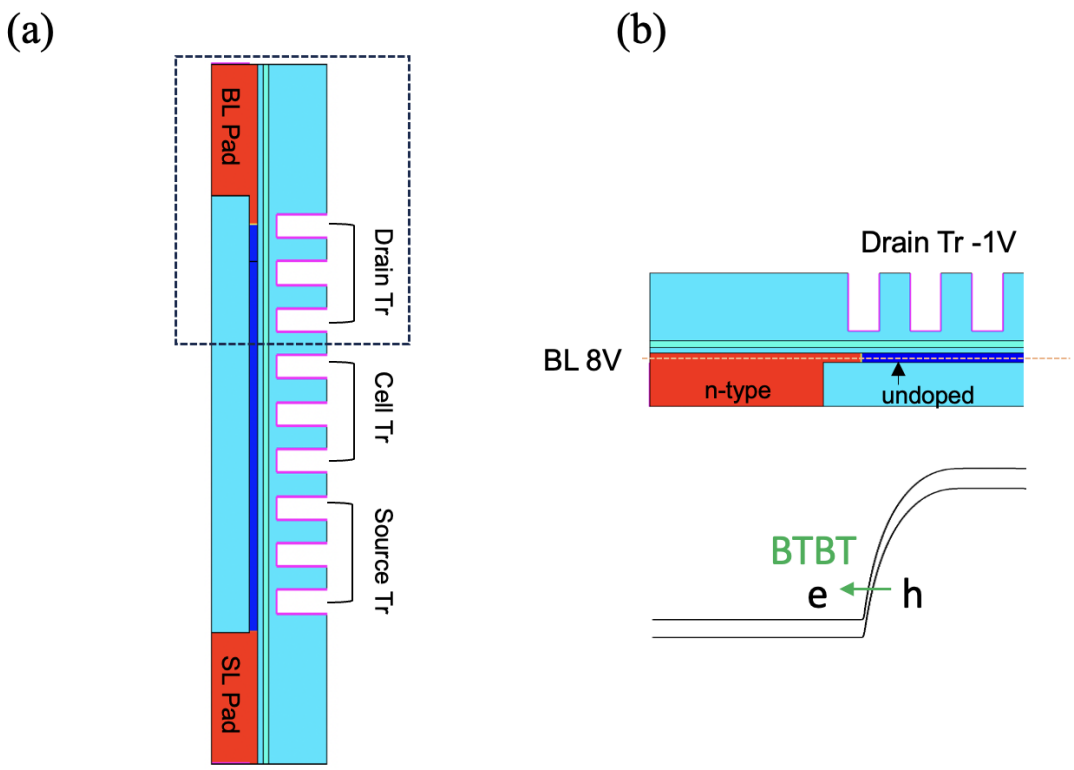


Figure 5.3: (a) Simplified 3D NAND string. (b) Energy band diagram along the channel near bit line (BL) and drain transistor during GIDL operation.

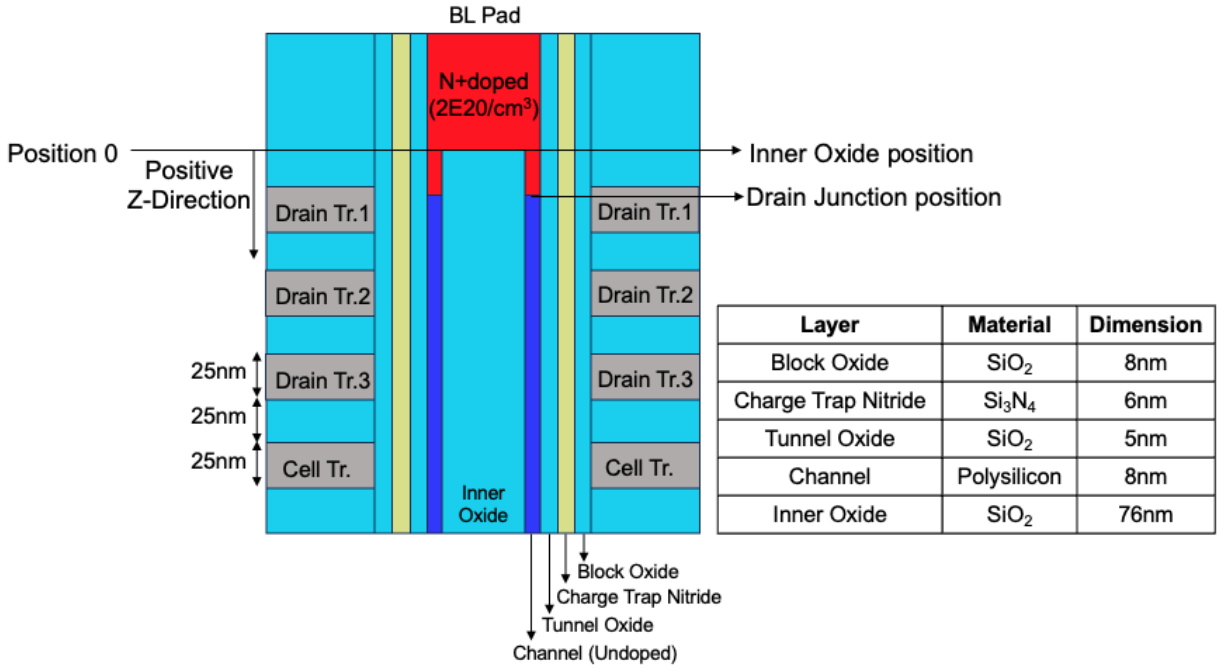
Compared to bulk erase, GIDL erase is preferred because it simplifies the string structure by eliminating the need for an additional p-type well. However, as the number of vertically stacked cells increases, the length of the polycrystalline-silicon (poly-Si) channel layer shared by the cells in a string also increases, which can result in slower erase operation [64]. The longer the channel layer, the more gradually the channel potential is boosted with the same amount of GIDL, and the slower the erase speed. Therefore, it is of great importance to enhance GIDL current to maintain fast erase speed as more cells are stacked in a 3D NAND string.

In this study, a silicon-germanium/silicon (SiGe/Si) heterojunction design for the drain transistor is proposed to enhance GIDL and hence erase speed, leveraging the smaller energy band gap of SiGe compared to Si to increase band-to-band tunneling current. TCAD simulations (Sentaurus from Synopsys [65]) are used to confirm the effectiveness of the SiGe/Si heterojunction for improving erase speed, and to optimize the drain transistor design while preserving inhibit mode operation.

5.2 3D NAND Flash Memory Simulation and Modeling

Simulation Structure

Figure 5.4 is a schematic cross-section of the uppermost portion of a 3D NAND string, showing the drain transistor (comprising three transistors in series, with their gates electrically tied together) and topmost memory cell, used for TCAD simulation study in this work. The undoped poly-Si channel region is indicated in dark blue, and the heavily doped n-type poly-Si drain region is indicated in red. The dimensions of each layers used for simulation are also include in the figure.



Layer	Material	Dimension
Block Oxide	SiO_2	8nm
Charge Trap Nitride	Si_3N_4	6nm
Tunnel Oxide	SiO_2	5nm
Channel	Polysilicon	8nm
Inner Oxide	SiO_2	76nm

Figure 5.4: Schematic cross-section of the uppermost portion of the 3D NAND string structure used in this simulation-based study.

GIDL operation modeling

To model the reverse-biased drain junction leakage during GIDL erase, both the Shockley–Read–Hall (SRH) model, which accounts for trap-assisted tunneling (TAT), and the band-to-band tunneling (BTBT) model, which accounts for direct band-to-band tunneling,

were employed. For GIDL operation, which measures the leakage current and serves as a key parameter determining the erase speed, the GIDL voltage (V_{gidl}) is defined as the voltage difference between the bit line (BL, connected to the drain region) and the gate of the drain transistor. Larger V_{gidl} induces steeper energy-band bending in the reverse-biased drain junction depletion region, resulting in more leakage current. Figure 5.5 shows simulations of the drain leakage current as a function of the BL voltage (V_{bl}), with the drain transistor gate biased at -1 V and the memory cell gate biased at 0 V. Both leakage mechanisms contribute to GIDL current. However, band-to-band tunneling (BTBT) is the dominant mechanism compared to Shockley–Read–Hall (SRH) leakage [66].

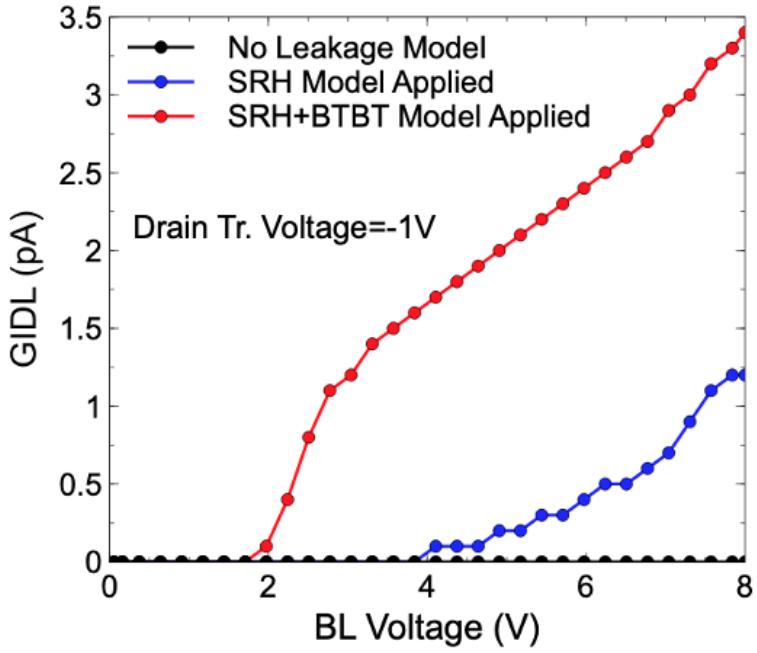


Figure 5.5: Simulated drain leakage current as a function of the BL voltage, depending on different leakage models. BTBT primarily contributes to GIDL.

Inhibit operation modeling

The drain transistors are used not only to generate GIDL during the erase operation but also to electrically isolate the string of memory cells in the inhibit mode of operation. In a 3D NAND flash memory cell array, each gate layer is shared with many strings within a block. Figure 5.6 shows a simplified illustration of two strings, represented by yellow lines. The left and right strings share the bit line (BL, shown in blue) and the cell transistor gates (shown in purple), while the drain transistor gates (shown in green) are independently controlled.

To program a memory cell in the left string at the center gate (marked as a red star), a high positive bias is applied to the selected gate. However, since the gate is shared by both left and right strings, the corresponding memory cell in the right string (marked as a yellow star) would be programmed unintentionally. To separate prevent this, the left string should be in program mode (i.e., with the drain transistor turned on), while the right-unselected strings should be in inhibit mode (i.e., with the drain transistor turned off).

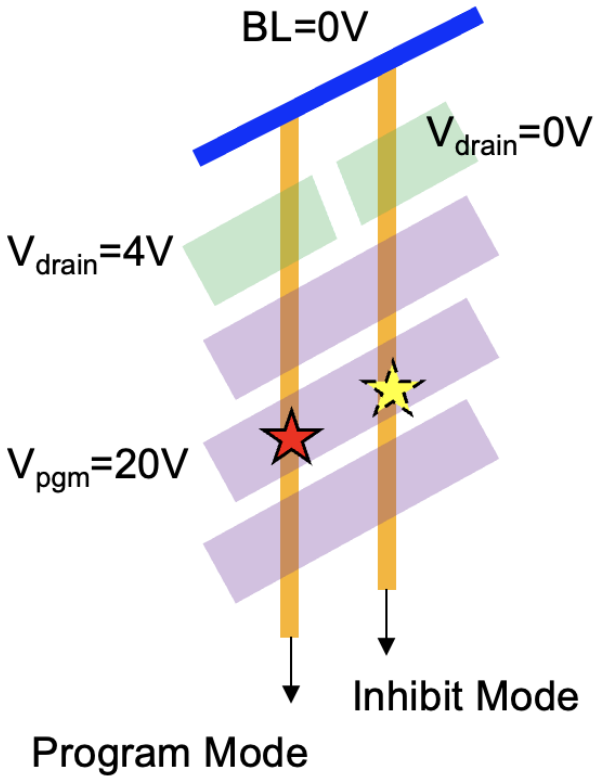
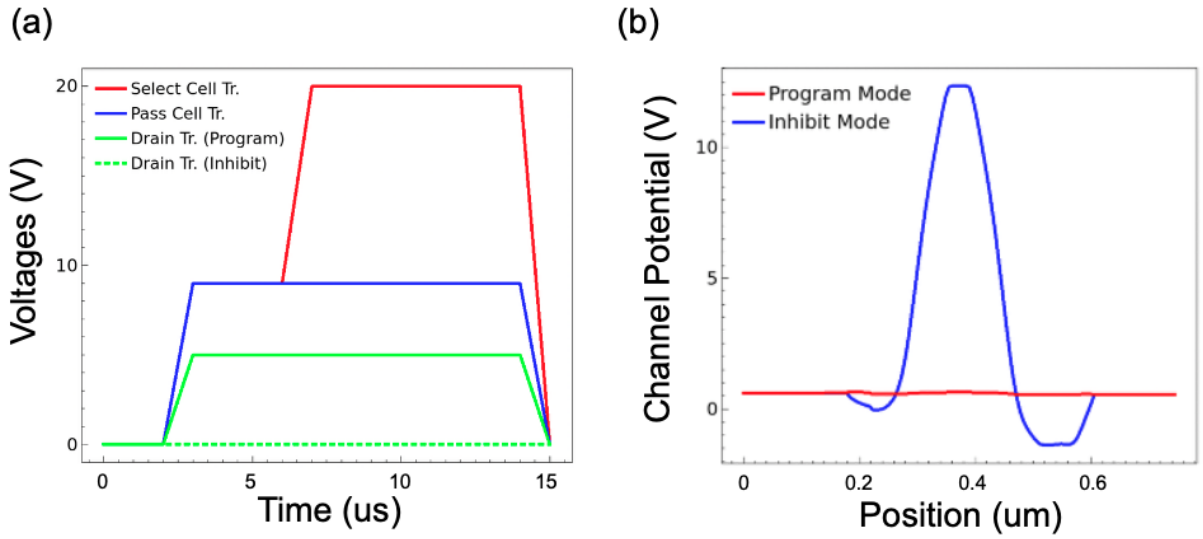


Figure 5.6: Simplified illustration of two strings, where a memory cell in the left string (red star) is programmed, while a memory cell in the right string (yellow star) is inhibited.

Figure 5.7 (a) shows the voltage waveforms for program and inhibit operation simulations. Figure 5.7 (b) compares the channel potential distributions for a string in program mode and in inhibit mode. In program mode, since the drain transistor is turned on, the channel region is grounded by the BL voltage. In contrast, in inhibit mode, the drain transistor turned off, electrically isolating the channel region from the BL. As a result, the applied program gate voltage (V_{pmg}) and applied pass gate voltage (V_{pass}) raise the potential of the

associated channel region, preventing a sufficiently large gate-to-channel voltage needed for programming operation.

The drain transistor must have low off-state leakage in inhibit mode [67] to ensure effective channel potential boosting, while it should generate high gate induced drain leakage (GIDL) in erase mode, which leads to a design tradeoff. Therefore, it is important to assess the impact of drain transistor design modifications on both the GIDL current during erase and the leakage current during inhibit operation.



Biasing Pad	Program Mode	Inhibit Mode
BL	0V	0V
Drain Tr.	4V	0V
Select Cell Tr.	V_{pgm} (20V)	V_{pgm} (20V)
Pass Cell Tr.	V_{pass} (9V)	V_{pass} (9V)
Source Tr.	0V	0V
SL	0V	0V

Figure 5.7: (a) Voltage waveforms for program and inhibit operations. The values of peak voltage applied to each electrode are indicated in the table. (b) Simulated channel potential distributions for a string in program mode (red line) and inhibit mode (blue curve). The channel region is grounded by the BL in program mode because the drain transistor is turned on. In inhibit mode, the channel region potential is elevated by V_{pmg} and V_{pass} because the drain transistor is turned off, disconnecting the channel from the BL.

5.3 SiGe/Si Heterojunction Drain Transistor for GIDL Enhancement

In this section, the efficacy of a SiGe/Si heterojunction drain transistor is evaluated in comparison with other approaches for increasing GIDL. Additionally, the correlation between increased GIDL and faster cell erase operation is studied.

The poly-Si channel region comprises crystalline grains and intermediary grain-boundary regions with dangling and strained bonds that serve as electron and hole trap sites with energy levels within the silicon energy bandgap, which can assist electron and hole tunneling [68]. Therefore, one approach to augmenting GIDL is to place a grain boundary within the aforementioned band-bending region. Figure 5.8 (a) shows that GIDL depends on the position of the grain boundary and that it indeed is maximized when the grain boundary is located within the high-field depletion region of the drain-channel junction. An increase in GIDL by a factor of 7 occurs when the grain boundary is located at 22 nm depth, but it diminishes dramatically with just 1 nm change in position. The random nature of grain boundary formation during the poly-Si deposition process makes it practically impossible to precisely control their position with respect to the doping junction.

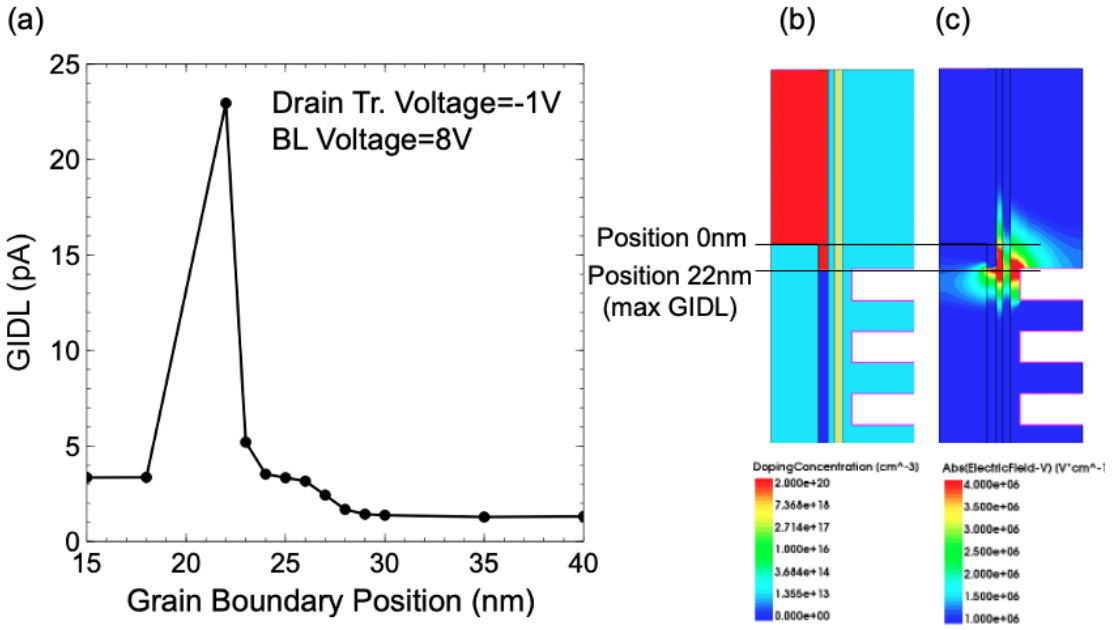


Figure 5.8: Erase mode: (a) Simulated GIDL as a function of polysilicon grain boundary position. (b) Illustration of the grain boundary position corresponding to peak GIDL. (c) Electric field contour plot. The electric field peaks at the drain (doping) junction; therefore GIDL enhancement by trap-assisted tunneling is maximized by placing the grain boundary at this location.

Another way to increase GIDL is to extend the heavily doped drain region into the thinner poly-Si region [69]. Figure 5.9 shows that a long drain junction results in higher peak electric field strength (Figure 5.9 (c)) and hence a higher BTBT rate (Figure 5.9 (d)) resulting in enhanced GIDL (Figure 5.9 (a)). However, this benefit comes at the cost of deteriorated off-state leakage current in inhibit mode, since a longer drain corresponds to shorter effective channel length of the drain transistor.

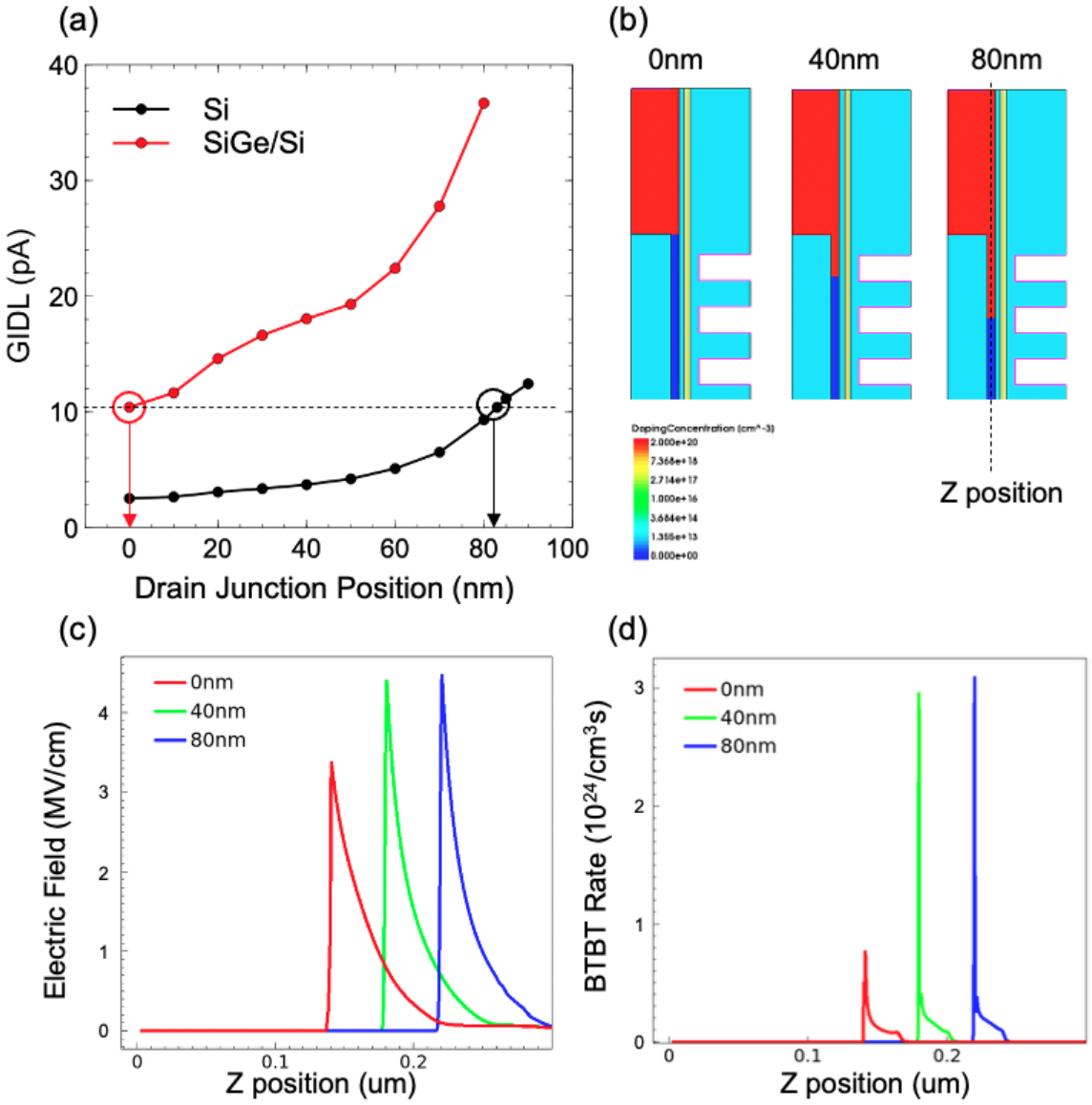


Figure 5.9: (a) Simulated GIDL dependence on drain junction position for Si (black curve) and SiGe/Si heterojunction (red curve) drain transistor. (b) Illustration of different drain junction positions. (c) Comparison of electric field profiles and (d) BTBT rate profiles for the different drain junction positions.

As an alternative to increasing the drain length, the use of SiGe as the heavily doped n-type drain material as well as a portion of the undoped thinner semiconductor region is proposed herein, as illustrated in Figure 5.10 (a). Note that the drain junction depletion

region resides mostly within the undoped side of the junction; this is why the smaller-band-gap SiGe material must extend into the undoped portion of the semiconductor layer for enhanced GIDL. At the same time, since a small-band-gap channel region can result in higher off-state (inhibit mode) leakage current [70], the SiGe should not extend too far into the channel region of the drain transistor; this is discussed in the following section. Herein the SiGe/Si heterojunction is fixed to be 20 nm below the drain (doping) junction. Figure 5.10 (c) compares the energy-band diagrams and BTBT rates for the conventional Si and new SiGe/Si heterojunction drain transistor designs, indicating that GIDL in erase mode is much higher for the SiGe/Si heterojunction drain transistor design. Therefore the SiGe/Si heterojunction drain transistor design eliminates the need for a long drain region to enhance GIDL in erase mode at a tradeoff of higher inhibit-mode leakage current.

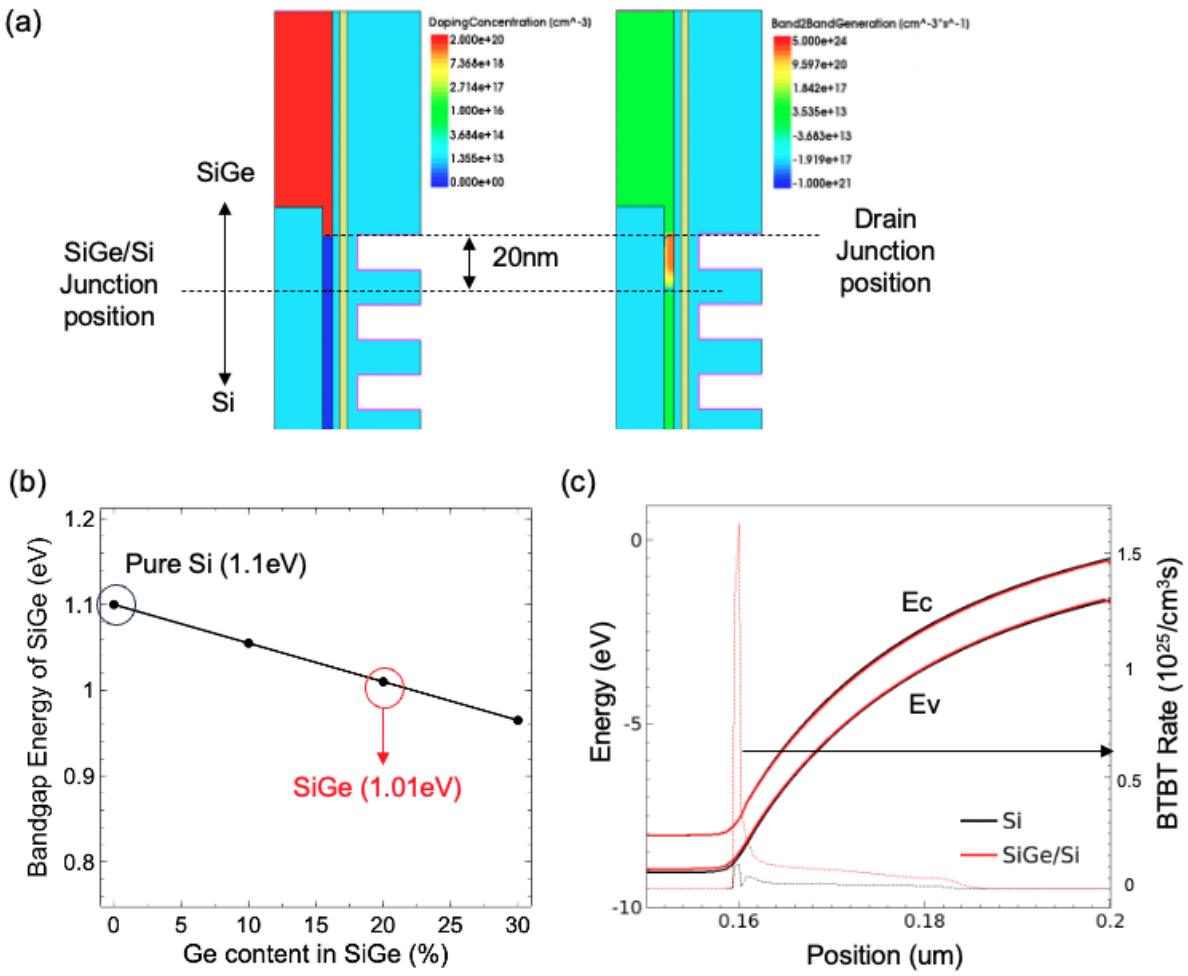


Figure 5.10: (a) Illustration of drain transistor (left) with the SiGe/Si heterojunction located 20 nm below the doping junction to maximize BTBT (right figure) occurring inside the SiGe. (b) Values of SiGe band-gap energy, dependent on Ge content, used in this study. (c) Comparison of energy band diagrams and BTBT rate for Si and SiGe (20 % Ge)/Si heterojunction drain transistor designs.

To further elucidate the benefit of the SiGe/Si heterojunction drain transistor, Figure 5.11 shows the relationship between GIDL current and the erased memory cell V_T for different drain junction lengths and Ge contents. For the same GIDL current (compare the black circle and red circle), the SiGe/Si heterojunction drain transistor achieves an erased V_T approximately 0.35 V lower than that of the conventional Si drain transistor, demonstrating faster erase operation.

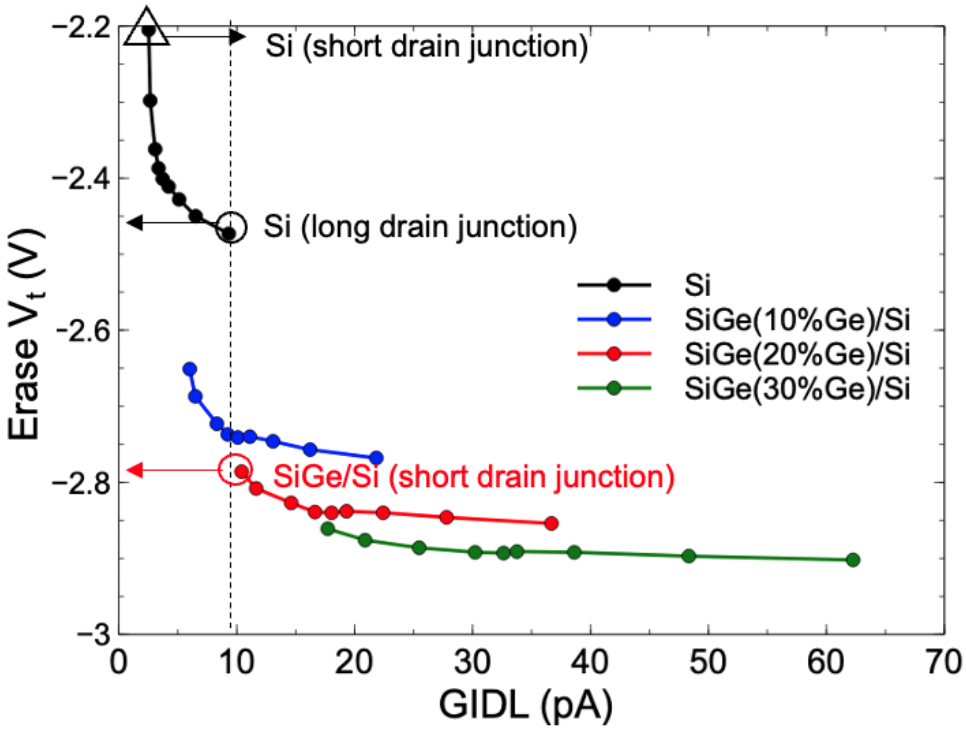


Figure 5.11: Simulated V_T value for an erased memory cell, as a function of the GIDL current, for various drain transistor designs. Even for the same value of GIDL, the SiGe/Si heterojunction drain transistor provides for greater erase speed (compare black circle and red circle).

Figure 5.12 shows the two drain transistor designs selected for comparison: the Si long-drain transistor and the SiGe/Si heterojunction drain transistor, both exhibiting the same GIDL current of 10 pA (corresponding to the black circle and red circle in Figure 5.11). These structures are compared to examine how the channel potential and trapped charge distributions in the memory cell evolve during the erase operation.

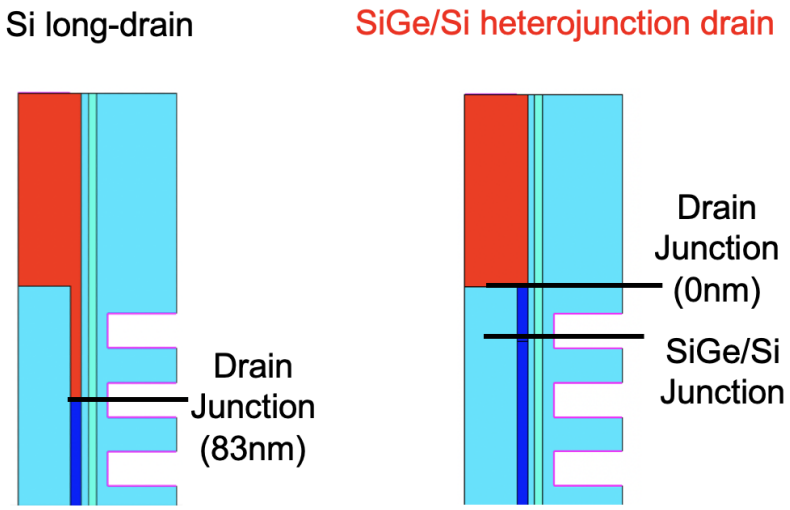


Figure 5.12: Two drain transistor designs that have the same GIDL of 10 pA: Si long-drain vs. SiGe/Si heterojunction drain

Figure 5.13 (a) shows the waveforms applied to the drain (BL), drain-transistor gate, and cell-transistor gate during the erase operation, defining five distinct time points. Figure 5.13 (b) compares the channel potential distributions for the Si long-drain vs. the SiGe/Si heterojunction drain transistor designs at each of the five points in time. It can be seen that the channel potential in the cell transistors increases more rapidly for the case of the SiGe/Si heterojunction drain transistor (red lines). As a result, the voltage drop across the ONO dielectric stack is greater so that electrons detrap from the nitride layer more quickly in the case of the SiGe/Si heterojunction drain transistor design, as shown in Figure 5.13 (c). Thus, the erased memory cell V_T is lower (more negative) for the SiGe/Si heterojunction drain transistor design.

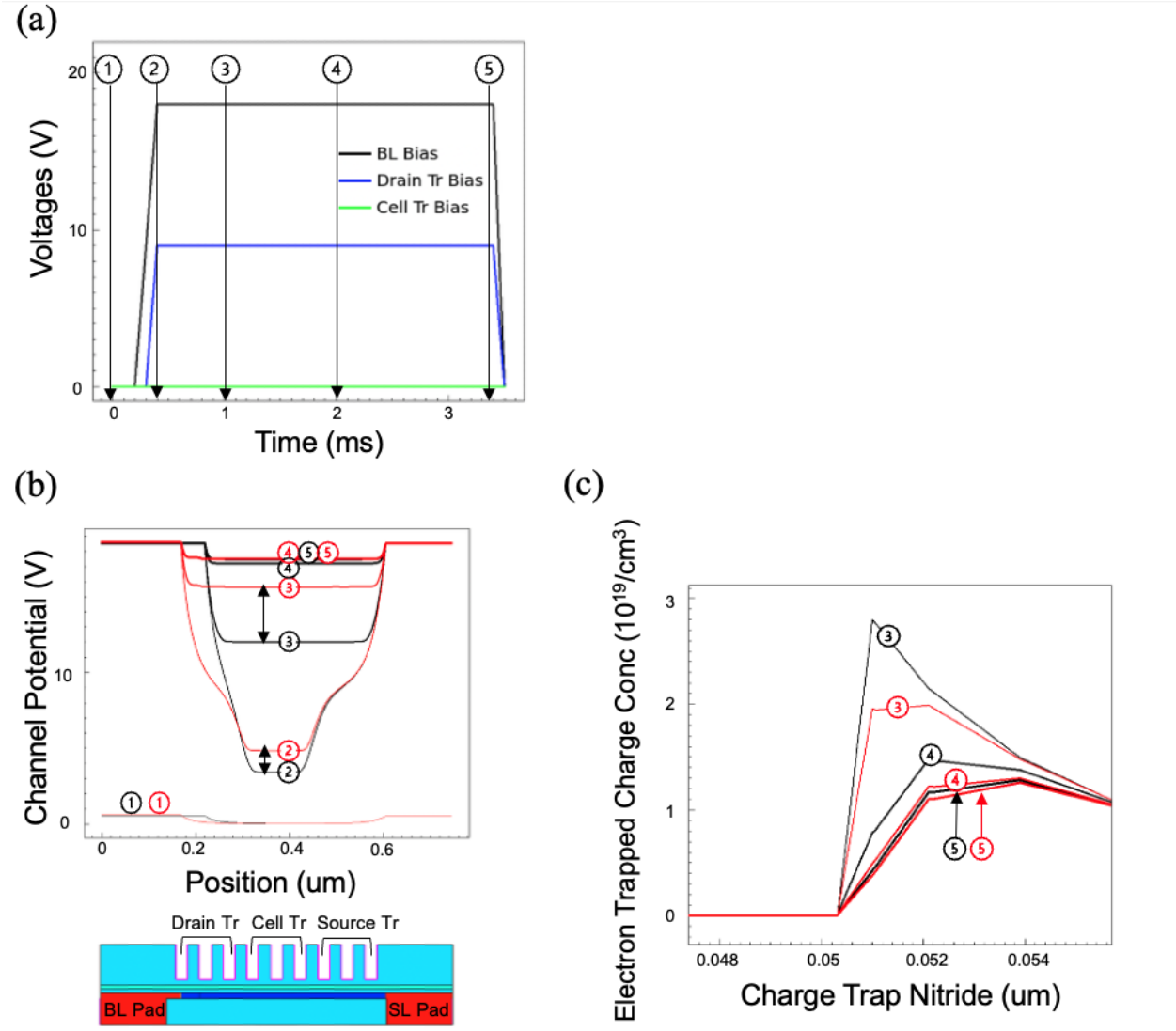


Figure 5.13: (a) Waveforms for the voltage signals applied to BL, drain transistor gate and memory cell transistor gates during erase operation, defining five points in time. (b) Channel potential distributions at each of the five points in time. The SiGe/Si heterojunction drain transistor increases channel potential faster than the Si long-drain transistor. (c) Electrons left in CTN at each of the five points in time. The SiGe/Si heterojunction drain transistor provides for faster erase (fewer remaining electrons stored in the CTN) due to more rapidly increased channel potential.

The faster increase in channel potential for the SiGe/Si heterojunction transistor can be explained by the shape of the GIDL characteristics shown in Figure 5.14. Although both

the Si long-drain transistor and the SiGe/Si heterojunction transistor exhibit the same final GIDL current of approximately 10 pA at high BL bias of 8 V, the SiGe/Si heterojunction produces a higher GIDL current in the lower BL bias region. During erase operation, the BL bias is gradually increased from a low voltage rather than being instantaneously applied at its maximum value. Consequently, the cumulative GIDL-generated hole injection is greater for the SiGe/Si heterojunction during the initial stage of erase. This larger early stage hole injection accelerates the rise in channel potential, leading to a faster erase operation.

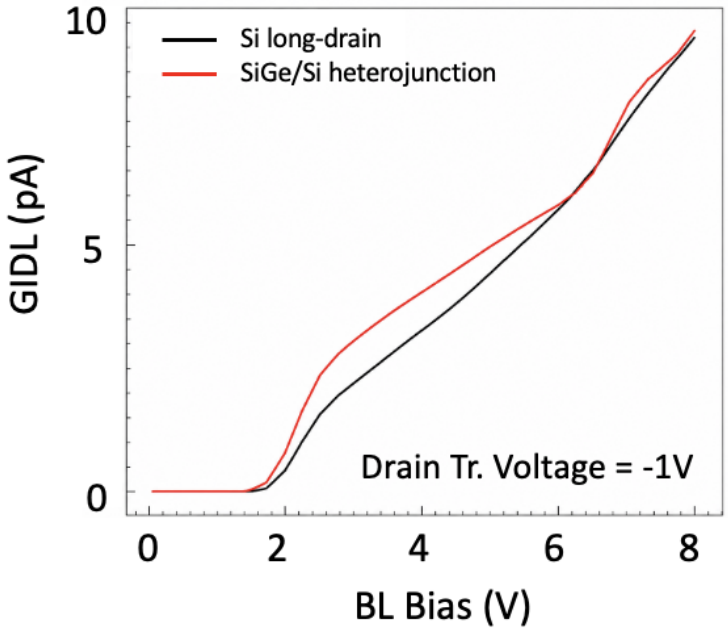


Figure 5.14: GIDL current as a function of BL bias for the Si long-drain and SiGe/Si heterojunction drain transistors.

Lastly, Figure 5.15 illustrates the process steps to form a SiGe heterojunction and the doped drain region: (a) chemical-mechanical polishing (CMP) is used to planarize the surfaces of the ONO, poly-Si and inner oxide layers; (b) the inner oxide layer is recessed; (c) the poly-Si layer is recessed; (d) undoped SiGe is selectively deposited over the poly-Si using Dichlorosilane (DCS); (e) doped SiGe is conformally deposited and CMP is used to planarize the surface. From a process integration perspective, the proposed approach is practical because it requires only one additional etch step (c) and one additional undoped SiGe deposition step (d) compared with the conventional drain transistor fabrication process.

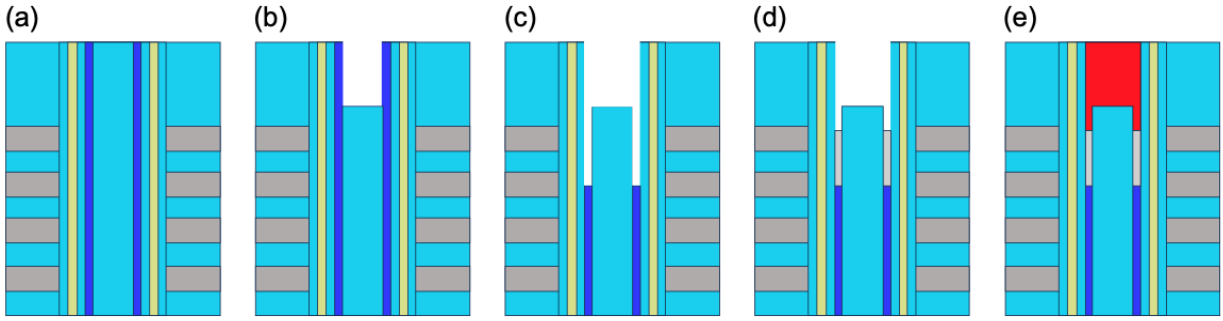


Figure 5.15: Process steps to form a SiGe/Si heterojunction and the doped drain region. (a) Planarization of the string through CMP. (b) Inner oxide is recessed. (c) Polysilicon channel layer is recessed. (d) Undoped SiGe is selectively deposited over the polysilicon. (e) N+ doped SiGe is deposited and planarized.

5.4 SiGe/Si Heterojunction Drain Transistor Design Optimization for Inhibit Mode Operation

In this section, inhibit-mode operation is studied in more detail. Figure 5.15 shows how the memory cell V_T during inhibit mode operation depends on the location of the drain junction, for three different semiconductor materials used in the drain transistor: Si, SiGe/Si heterojunction, and SiGe. The cell V_T is initially set at -2 V; if the cell is disturbed in inhibit mode, then the cell V_T increases (i.e., becomes less negative) due to unintended program. For the Si and SiGe/Si heterojunction drain transistor designs with drain length below 60 nm, the cell Vt values are nearly identical, indicating that partial incorporation of SiGe does not adversely affect inhibit mode operation. If the drain length exceeds 60 nm, the cell Vt is significantly degraded, slightly more so for the SiGe/Si heterojunction drain transistor design. If SiGe is entirely incorporated throughout the drain transistor, however, then the memory cell V_T is significantly disturbed during inhibit mode operation, increasingly so with the length of the drain region. This is due to the higher off-state leakage current associated with a smaller band-gap energy semiconductor and/or shorter channel length transistor.

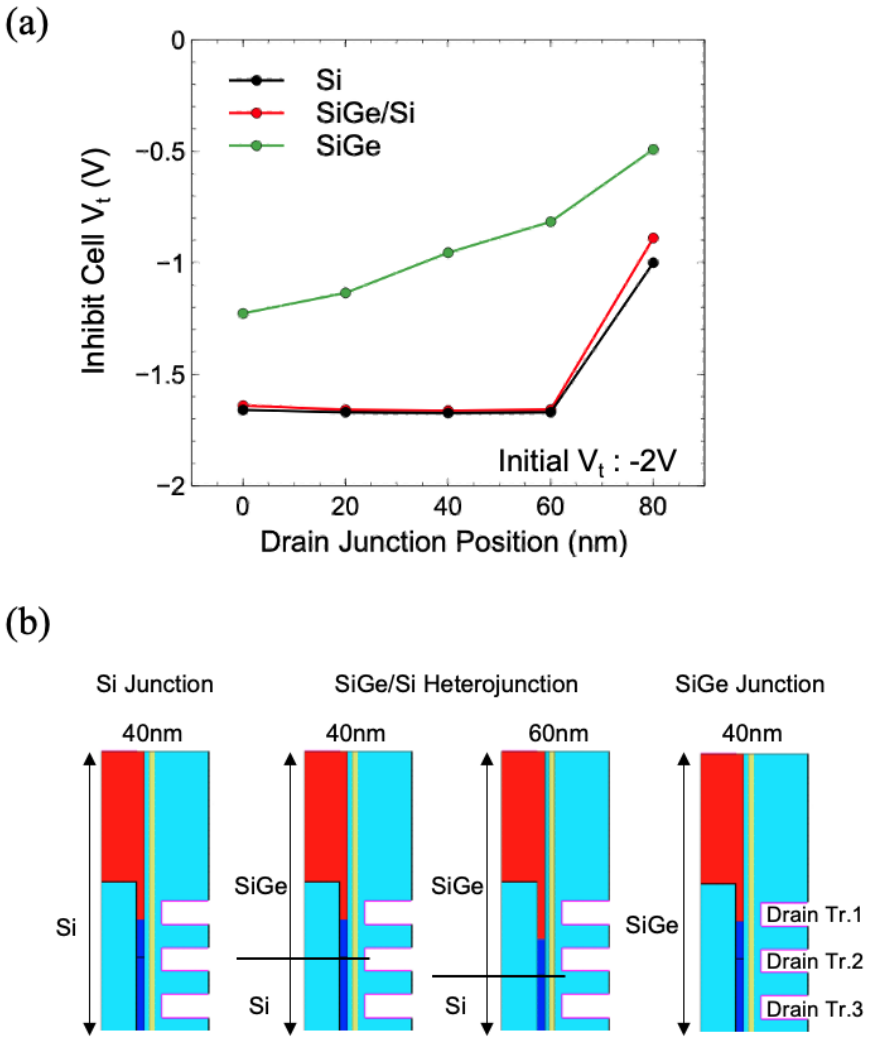


Figure 5.16: Inhibit mode: (a) Memory cell V_T dependence on the drain junction position, for Si, SiGe/Si heterojunction and SiGe drain transistors. The Si and SiGe/Si heterojunction drain transistors show similar cell Vt trend when the drain length is below 60 nm, while the SiGe drain transistor shows significantly deteriorated cell V_T. (b) Illustrations of Si (left), SiGe/Si heterojunction (center two figures) and SiGe (right) drain transistors. Note that the Si channel portion in the SiGe/Si heterojunction drain transistor is reduced as the drain length increases since the distance between the drain junction and SiGe/Si heterojunction is fixed at 20 nm.

5.5 Conclusion

A practical SiGe/Si heterojunction drain transistor design is proposed to enhance GIDL for faster 3D NAND erase operation, leveraging SiGe's smaller energy band gap. With only 20 % Ge in SiGe, GIDL is projected to increase by 5× compared to the conventional Si drain transistor design. Even for the same peak GIDL level, the SiGe/Si heterojunction drain transistor design is projected to provide for faster erase operation. Optimal placement of the SiGe/Si heterojunction can ensure that there is no negative impact on inhibit mode operation.

Chapter 6

Conclusion

This dissertation investigated approaches to improve the reliability and performance of advanced CMOS and memory devices for future computing systems. As artificial intelligence continues to increase the demand for computing capability, semiconductor technologies must continue to improve transistor performance, power efficiency, reliability, and memory cell density. Addressing these challenges requires innovations in materials, device structures, and fabrication technologies. This dissertation presented four studies covering defect physics, high-k/metal-gate engineering, work function modulation through Oxygen insertion (OI) technology, and advanced 3D NAND Flash design.

Chapter 2 examined random telegraph noise (RTN) in advanced MOSFETs under cryogenic operation through electrical characterization and TCAD simulation. The study clarified the dependence of RTN behavior on temperature, gate bias, and drain bias through a modified Shockley–Read–Hall (SRH) recombination model. The analysis further demonstrated how transistor geometry influences RTN characteristics. These findings improve the understanding of defect behavior in advanced CMOS technologies and provide valuable insight for the design and optimization of future cryogenic computing.

Chapter 3 focused on work function engineering through the incorporation of an ultrathin Al layer within the HKMG stack. Electrical characterization combined with chemical analysis revealed that oxygen scavenging by the ultrathin Al layer reduces the oxygen concentration in the TiON layer, shifting its composition toward TiN and thereby reducing the effective gate work function. The study also identified the tradeoff between work function modulation and dielectric reliability associated with excessive Al incorporation into the gate dielectric. These findings clarify the physical mechanisms governing work function engineering through ultrathin Al incorporation in HKMG technology and the associated reliability tradeoffs.

Chapter 4 investigated oxygen insertion (OI) technology as an alternative approach for work function engineering. The OI layer modulates the effective work function by suppressing intermixing within the HKMG stack while simultaneously improving carrier mobility.

Electrical characterization together with SIMS and EELS analyses showed that the OI layer suppresses oxygen migration toward the metallic gate, thereby reducing TiON formation and preserving a higher fraction of TiN. Unlike conventional work function engineering, which relies on increasingly complex HKMG stacks, the OI approach enables threshold voltage control by modifying the silicon substrate, providing a potentially scalable approach for future CMOS technologies.

Chapter 5 proposed a SiGe/Si heterojunction drain transistor to improve erase operation in 3D NAND flash memory. TCAD simulation demonstrated that introducing a SiGe region near the drain enhances band-to-band tunneling and increases gate induced drain leakage, enabling a more rapid increase in channel potential during erase operation. The proposed structure was further shown to maintain inhibit mode characteristics while improving erase speed. These results demonstrate that heterojunction engineering provides an effective approach for improving erase performance without compromising inhibit mode operation.

Although each study addressed a different technological challenge, they collectively highlight the critical role of material engineering and device physics in advancing semiconductor technology. As device dimensions continue to shrink and fabrication complexity increases, phenomena such as defect behavior, oxygen transport, work function control, and band engineering will become increasingly important. The findings presented in this dissertation provide a foundation for the continued development of high-performance and reliable semiconductor devices.

One important direction for future research is the experimental validation of OI technology in advanced gate-all-around CMOS devices, where increasingly narrow nanosheet spacing makes conventional work function engineering more difficult. Moreover, continued advances in materials engineering, process optimization, and device design will remain essential for developing future CMOS and memory technologies that support next-generation computing systems.

Bibliography

[1] Ashish Vaswani et al. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 5998–6008.

[2] Nicola Jones. "How to stop data centres from gobbling up the world's electricity". In: *Nature* 561 (2018), pp. 163–166. DOI: 10.1038/d41586-018-06610-y.

[3] A. Mehonic and A. J. Kenyon. "Brain-inspired computing needs a master plan". In: *Nature* 604 (2022), pp. 255–259. DOI: 10.1038/s41586-021-04362-w.

[4] Asif Ali Khan et al. "The Landscape of Compute-near-memory and Compute-in-memory: A Research and Commercial Overview". In: *arXiv preprint arXiv:2401.14428* (2024). DOI: 10.48550/arXiv.2401.14428.

[5] Abu Sebastian et al. "Memory Devices and Applications for In-Memory Computing". In: *Nature Nanotechnology* 15.7 (2020), pp. 529–544. DOI: 10.1038/s41565-020-0655-z.

[6] Hongshin Jun et al. "HBM (High Bandwidth Memory) DRAM Technology and Architecture". In: *2017 IEEE International Memory Workshop (IMW)*. 2017, pp. 1–4. DOI: 10.1109/IMW.2017.7939084.

[7] Sayeef Salahuddin, Kai Ni, and Suman Datta. "The era of hyper-scaling in electronics". In: *Nature Electronics* 1 (2018), pp. 442–450. DOI: 10.1038/s41928-018-0117-x.

[8] Robert H. Dennard et al. "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions". In: *IEEE Journal of Solid-State Circuits* 9.5 (1974), pp. 256–268. DOI: 10.1109/JSSC.1974.1050511.

[9] David J. Frank et al. "Device Scaling Limits of Si MOSFETs and Their Application Dependencies". In: *Proceedings of the IEEE* 89.3 (2001), pp. 259–288. DOI: 10.1109/5.915374.

[10] Chenming C. Hu. *Modern Semiconductor Devices for Integrated Circuits*. Upper Saddle River, NJ: Prentice Hall, 2010. ISBN: 9780136085256.

[11] Robert H. Dennard et al. "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions". In: *IEEE Journal of Solid-State Circuits* 9.5 (1974), pp. 256–268. DOI: 10.1109/JSSC.1974.1050511.

[12] K. Mistry et al. “A 45 nm Logic Technology with High-k + Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193 nm Dry Patterning, and 100% Pb-Free Packaging”. In: *2007 IEEE International Electron Devices Meeting (IEDM)*. 2007, pp. 247–250. DOI: 10.1109/IEDM.2007.4418914.

[13] H. Wagner et al. “The Work Function Difference of the MOS-System with Aluminium Field Plates and Polycrystalline Silicon Field Plates”. In: *Solid-State Electronics* 17.8 (1974), pp. 769–775. DOI: 10.1016/0038-1101(74)90023-9.

[14] D. Hisamoto et al. “FinFET—A Self-Aligned Double-Gate MOSFET Scalable to 20 nm”. In: *IEEE Transactions on Electron Devices* 47.12 (2000), pp. 2320–2325. DOI: 10.1109/16.887014.

[15] M. Loubet et al. “Stacked Nanosheet Gate-All-Around Transistor to Enable Scaling Beyond FinFET”. In: *2017 IEEE Symposium on VLSI Technology*. 2017, T230–T231. DOI: 10.23919/VLSIT.2017.7998183.

[16] Kin P. Cheung. “On the 60 mV/dec @300 K Limit for MOSFET Subthreshold Swing”. In: *2010 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*. 2010.

[17] M. J. Deen and F. Pascal. “Cryogenic Operation of CMOS-Based Microsystems and Computers”. In: *Microprocessors and Microsystems* 13.4 (1989), pp. 245–253. DOI: 10.1016/0141-9331(89)90062-8.

[18] Kiseok Lee et al. “Monolithically Stackable 1T1C 3-D DRAM: A Technical Survey”. In: *IEEE Transactions on Electron Devices* 73.3 (2026), pp. 1092–1112. DOI: 10.1109/TED.2026.3659106.

[19] Feng-Min Lee et al. “Bit-Cost-Scalable 3D DRAM Architecture and Unit Cell First Demonstrated with Integrated Gate-Around and Channel-Around IGZO FETs”. In: *2024 IEEE Symposium on VLSI Technology and Circuits*. 2024, pp. 1–2. DOI: 10.1109/VLSITechnologyandCir46783.2024.10631386.

[20] Charles Kuo, Tsu-Jae King, and Chenming Hu. “A Capacitorless Double-Gate DRAM Cell Design for High Density Applications”. In: *International Electron Devices Meeting (IEDM)*. 2002, pp. 843–846. DOI: 10.1109/IEDM.2002.1175955.

[21] Hongtao Zhong, Zijie Zheng, Leming Jiao, et al. “First Demonstration of AFeFET-Based Capacitor-Less eDRAM Computing-in-Memory Featuring 4.84 Mb/mm^2 High Memory Density, 10^5 s Long Retention Time, and $> 10^{10}$ High Endurance”. In: *2024 IEEE International Electron Devices Meeting (IEDM)*. 2024. DOI: 10.1109/IEDM50854.2024.10873490.

[22] Eun-Seok Choi and Sung-Kye Park. “Device Considerations for High Density and Highly Reliable 3D NAND Flash Cell in Near Future”. In: *2012 IEEE International Electron Devices Meeting (IEDM)*. 2012, pp. 9.4.1–9.4.4. DOI: 10.1109/IEDM.2012.6479011.

[23] Youngjun Park and Wonbo Shim. "Incremental Pulse-Width Erase (IPWE) Scheme for Fast and Variation-Tolerant GIDL Erase of 3D NAND Flash". In: *Micromachines* 17.4 (2026), p. 399. DOI: 10.3390/mi17040399.

[24] Rakshith Saligram, Suman Datta, and Arijit Raychowdhury. "Cryogenic CMOS as an Enabler for Low Power Dynamic Logic". In: *2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. 2023. DOI: 10.1109/ISLPED58423.2023.10244724.

[25] Kiyoshi Takeuchi. "Impact of Discrete-Charge-Induced Variability on Scaled MOS Devices". In: *IEICE Transactions on Electronics* E95-C.4 (2012), pp. 414–420. DOI: 10.1587/transele.E95.C.414.

[26] Salvatore Maria Amoroso et al. "3D Dynamic RTN Simulation of a 25 nm MOSFET: The Importance of Variability in Reliability Evaluation of Decananometer Devices". In: *2012 15th International Workshop on Computational Electronics (IWCE)*. 2012. DOI: 10.1109/IWCE.2012.6242823.

[27] H. D. Dixit et al. "Silent Data Corruptions at Scale". In: *arXiv preprint arXiv:2102.11245* (2021). DOI: 10.48550/arXiv.2102.11245.

[28] Kevin Gibb. *Samsung Delivers Its 14nm*. eeNews Europe. Online. Apr. 2015. URL: https://www.eenewseurope.com/en/samsung-delivers-its-14nm.

[29] HeeBong Yang et al. "Random Telegraph Noise of a 28-nm Cryogenic MOSFET in the Coulomb Blockade Regime". In: *IEEE Electron Device Letters* 42.1 (2021), pp. 5–8. DOI: 10.1109/LED.2021.3132964.

[30] R. Asanovski et al. "New Insights on the Excess 1/f Noise at Cryogenic Temperatures in 28 nm CMOS and Ge MOSFETs for Quantum Computing Applications". In: *2022 International Electron Devices Meeting (IEDM)*. 2022. DOI: 10.1109/IEDM45625.2022.10019388.

[31] Y. F. Lim et al. "Random Telegraph Signal Noise in Gate-All-Around Si-FinFET With Ultranarrow Body". In: *IEEE Electron Device Letters* 27.9 (2006), pp. 765–768. DOI: 10.1109/LED.2006.880640.

[32] Wen Fang et al. "Distinction Between Silicon and Oxide Traps Using Single-Trap Spectroscopy". In: *Physica Status Solidi A* 212.3 (2015). DOI: 10.1002/pssa.201400087.

[33] *Sentaurus Device User Guide*. Version T-2022.03. Synopsys. 2022.

[34] M. J. Kirton and M. J. Uren. "Noise in Solid-State Microstructures: A New Perspective on Individual Defects, Interface States and Low-Frequency (1/f) Noise". In: *Advances in Physics* 38.4 (1989), pp. 367–468. DOI: 10.1080/00018738900101122.

[35] Toshiharu Nagumo et al. "Statistical Characterization of Trap Position, Energy, Amplitude and Time Constants by RTN Measurement of Multiple Individual Traps". In: *2010 International Electron Devices Meeting (IEDM)*. 2010. DOI: 10.1109/IEDM.2010.5703437.

[36] Yichao Sun et al. “The Drain Bias Modulation Effect of Random Telegraph Noise in Gate-All-Around FETs for Cryogenic Applications”. In: *IEEE Electron Device Letters* 45.4 (2024), pp. 530–533. DOI: 10.1109/LED.2024.3362986.

[37] Z. M. Shi, J.-P. Miéville, and M. Dutoit. “Effect of Electron Heating on Electron Capture Cross Section in Very Small Metal-Oxide-Semiconductor Transistors”. In: *Applied Physics Letters* 62.18 (1993). DOI: 10.1063/1.109426.

[38] *International Roadmap for Devices and Systems (IRDS)*. Online. 2023. URL: https://irds.ieee.org/.

[39] D. James. “High-k/Metal Gates in Leading Edge Silicon Devices”. In: *2012 SEMI Advanced Semiconductor Manufacturing Conference*. Saratoga Springs, NY, USA, 2012, pp. 346–353. DOI: 10.1109/ASMC.2012.6212925.

[40] H. Arimura et al. “Shifter Materials and Stack Explorations for Vt Fine-Tunable Dual-Dipole Multi-Vt Gate Stacks Compatible with Low-Thermal-Budget CFET”. In: *2025 Symposium on VLSI Technology and Circuits*. Kyoto, Japan, 2025, pp. 1–3. DOI: 10.23919/VLSITechnologyandCir65189.2025.11075025.

[41] *ASM International Presentation*. SEMI. Online. URL: https://www.semi.org/sites/semi.org/files/2022-11/06_MichaelGivens_ASMInt.pdf.

[42] Feliciano Giustino, Paolo Umari, and Alfredo Pasquarello. “Dielectric Effect of a Thin SiO2 Interlayer at the Interface Between Silicon and High-k Oxides”. In: *Microelectronic Engineering* 72 (2004), pp. 299–303. DOI: 10.1016/j.mee.2004.01.011.

[43] K. Xiong et al. “Defect Energy Levels in High-Dielectric-Constant Gate Oxide”. In: *Applied Physics Letters* 87 (2005), p. 183505. DOI: 10.1063/1.2119425.

[44] E. Cartier et al. “Fundamental Understanding and Optimization of PBTI in nFETs with SiO2/HfO2 Gate Stack”. In: *2006 International Electron Devices Meeting*. San Francisco, CA, USA, 2006, pp. 1–4. DOI: 10.1109/IEDM.2006.346773.

[45] T. R. Oldham, A. J. Lelis, and F. B. McLean. “Spatial Dependence of Trapped Holes Determined from Tunneling Analysis and Measured Annealing”. In: *IEEE Transactions on Nuclear Science* 33.6 (Dec. 1986), pp. 1203–1209. DOI: 10.1109/TNS.1986.4334579.

[46] R. Yamada et al. “Analysis of Detrap Current Due to Oxide Traps to Improve Flash Memory Retention”. In: *2000 IEEE International Reliability Physics Symposium Proceedings*. San Jose, CA, USA, 2000, pp. 200–204. DOI: 10.1109/RELPHY.2000.843915.

[47] Y. C. Yeo, T.-J. King, and C. Hu. “Direct Tunneling Leakage Current and Scalability of Alternative Gate Dielectrics”. In: *Applied Physics Letters* 81.11 (Sept. 2002), pp. 2091–2093. DOI: 10.1063/1.1506941.

[48] Walter E. Dahlke and Sanjay Jain. "MOS Tunneling Rate and Interface State Capture Cross-Section". In: *The Physics and Chemistry of SiO2 and the Si-SiO2 Interface*. Ed. by C. R. Helms and B. E. Deal. Boston, MA: Springer. DOI: 10.1007/978-1-4899-0774-5_41.

[49] A. M. Sonnet et al. "Impact of Semiconductor and Interface-State Capacitance on Metal/High-k/GaAs Capacitance–Voltage Characteristics". In: *IEEE Transactions on Electron Devices* 57.10 (Oct. 2010), pp. 2599–2605. DOI: 10.1109/TED.2010.2058853.

[50] R. G. Forbes. "Field Emission: New Theory for the Derivation of Emission Area from a Fowler–Nordheim Plot". In: *Journal of Vacuum Science & Technology B* 17.2 (Mar. 1999), pp. 526–533. DOI: 10.1116/1.590588.

[51] Yu-Lun Chueh et al. "RuO2 Nanowires and RuO2/TiO2 Core/Shell Nanowires: From Synthesis to Mechanical, Optical, Electrical, and Photoconductive Properties". In: *Advanced Materials* 19.1 (2007), pp. 143–149.

[52] M. Abbate et al. "Soft X-Ray Absorption Spectroscopy of Vanadium Oxides". In: *Journal of Electron Spectroscopy and Related Phenomena* 62 (1993), pp. 185–195.

[53] J. C. Sánchez-López, A. Caballero, and A. Fernández. "Characterisation of Passivated Aluminium Nanopowders: An XPS and TEM/EELS Study". In: *Journal of the European Ceramic Society* 18 (1998), pp. 1195–1200.

[54] Kenzo Manabe et al. "Comprehensive Analysis of Determining Factors of the Effective Work Function of TiAlC/TiN Metal Gates for Advanced Gate-All-Around CMOS Integration". In: *Japanese Journal of Applied Physics* 64.1 (Jan. 2025). DOI: 10.35848/1347-4065/ada9f8.

[55] J. A. Smith et al. "Simultaneous Carrier Transport Enhancement and Variability Reduction in Si MOSFETs by Insertion of Partial Monolayers of Oxygen". In: *2012 IEEE Silicon Nanoelectronics Workshop (SNW)*. 2012.

[56] N. Xu et al. "Electron Mobility Enhancement in (100) Oxygen-Inserted Silicon Channel". In: *Applied Physics Letters* 107.12 (2015), p. 123502. DOI: 10.1063/1.4931644.

[57] Jeffrey A. Smith et al. "Intermixing Reduction in Ultra-Thin Titanium Nitride/Hafnium Oxide Film Stacks Grown on Oxygen-Inserted Silicon and Associated Reduction of the Interface Charge Dipole". In: *Journal of Applied Physics* 130.18 (2021), p. 185303. DOI: 10.1063/5.0068002.

[58] Hideki Takeuchi et al. "Remote Control of Doping Profile, Silicon Interface, and Gate Dielectric Reliability via Oxygen Insertion into Silicon Channel". In: *2022 IEEE International Meeting for Future of Electron Devices, Kansai (IMFEDK)*. 2022, pp. 1–4. DOI: 10.1109/IMFEDK56875.2022.9975306.

[59] Vikram Magendar et al. "Electron and Hole Surface Roughness Scattering-Limited Mobilities in Oxygen-Inserted (OI) Si Channel". In: *Journal of Applied Physics* 139.12 (2026), p. 125705. DOI: 10.1063/5.0316691.

[60] Jeffrey A. Smith et al. "Intermixing Reduction in Ultra-Thin Titanium Nitride/Hafnium Oxide Film Stacks Grown on Oxygen-Inserted Silicon and Associated Reduction of the Interface Charge Dipole". In: *Journal of Applied Physics* 130.18 (2021), p. 185303. DOI: 10.1063/5.0068002.

[61] E. H. Nicollian and A. Goetzberger. "The Si–SiO_2 Interface – Electrical Properties as Determined by the Metal-Insulator-Silicon Conductance Technique". In: *Bell System Technical Journal* 46.6 (1967), pp. 1055–1133. DOI: 10.1002/j.1538-7305.1967.tb01712.x.

[62] Sz-Chian Liou et al. "Electronic Excitations in Monoclinic Hafnia as Studied by Spatially and Momentum-Resolved Electron Energy-Loss Spectroscopy and Ab Initio Density Functional Theory Calculations". In: *Physical Review Materials* 7.6 (2023), p. 065201. DOI: 10.1103/PhysRevMaterials.7.065201.

[63] D. A. Muller et al. "The Electronic Structure at the Atomic Scale of Ultrathin Gate Oxides". In: *Nature* 399.6738 (1999), pp. 758–761. DOI: 10.1038/21602.

[64] Akira Goda. "Recent Progress on 3D NAND Flash Technologies". In: *Electronics* (2021).

[65] Synopsys. *Sentaurus Device User Guide*. Version T-2022.03. Synopsys. 2022.

[66] Vishal A. Tiwari et al. "Analysis of Gate-Induced Drain Leakage Mechanisms in Silicon-Germanium Channel pFET". In: *IEEE Transactions on Electron Devices* (2014).

[67] Yu Zhang et al. "Leakage Characterization of Top Select Transistor for Program Disturbance Optimization in 3D NAND Flash". In: *Solid-State Electronics* (2018).

[68] W. J. Wu et al. "A Compact Model for Polysilicon TFT Leakage Current Including the Poole–Frenkel Effect". In: *IEEE Transactions on Electron Devices* (2007).

[69] Suhwan Lim et al. "Improvement of GIDL-Assisted Erase by Using Surrounded BL PAD Structure for VNAND". In: *IEEE International Memory Workshop (IMW)*. 2023.

[70] M. V. Fischetti and S. E. Laux. "Band Structure, Deformation Potentials, and Carrier Mobility in Strained Si, Ge, and SiGe Alloys". In: *Journal of Applied Physics* (1996).