ABSTRACT

Title of dissertation: DEVICE AND CIRCUIT LEVEL EMI INDUCED VULNERABILITY: MODELING AND EXPERIMENTS

Yumeng Cui, Doctor of Philosophy, 2021

Dissertation directed by: Professor Neil Goldsman
Department of Electrical and Computer Engineering

Electro-magnetic interference (EMI) commonly exists in electronic equipment containing semiconductor-based integrated circuits (ICs). Metal-oxide-semiconductor field-effect-transistors (MOSFETs) in the ICs may be disrupted under EMI conditions due to transient voltage-current surges, and their internal states may change undesirably. In this work, the vulnerabilities of silicon MOSFETs under EMI are studied at the device and the circuit levels, categorized as non-permanent upsets ("Soft Errors") and permanent damages ("Hard Failures").

The Soft Errors, such as temporary bit errors and waveform distortions, may happen or be intensified under EMI, as the transient disruptions activate unwanted and highly non-linear changes inside MOSFETs, such as impact ionization and Snapback. The system may be corrected from the erroneous state when the EMI condition is removed. We simulate planar silicon n-type MOSFETs at the device level to study the physical mechanisms leading to or complicate the short-term, signal-level Soft Errors. We experimentally tested commercially available MOSFET devices. Not included in regular MOSFET models, exponential-like current increases as the terminal voltage increases are observed and explained using the device-level knowledge. We develop a compact Soft Error model, compatible with circuit simulators using lumped (or compact-model) components and closed-

form expressions, such as SPICE, and calibrate it with our in-house experimental data using an in-house extraction technique based on the Genetic Algorithm. Example circuits are simulated using the extracted device model and under EMI-induced transient disruptions.

The EMI voltage-current disruptions may also lead to permanent Hard Failures that cannot be repaired without replacement. One type of Hard Failures, the MOSFET gate dielectric (or "oxide") breakdown, can result in input-output relation changes and additional thermal runaway. We have fabricated individual MOSFET devices at the FabLab at the University of Maryland NanoCenter. We experimentally stress-test the fabricated devices and observe the rapid, permanent oxide breakdown. Then, we simulate a nano-scale FinFET device with ultra-thin gate oxide at the device level. Then, we apply the knowledge from our experiments to the simulated FinFET, producing a gate oxide breakdown Hard Failure circuit model.

The proposed workflow enables the evaluation of EMI-induced vulnerabilities in circuit simulations before actual fabrication and experiments, which can help the early-stage prototyping process and reduce the development time.

DEVICE AND CIRCUIT LEVEL EMI INDUCED VULNERABILITY: MODELING AND EXPERIMENTS

by

Yumeng Cui

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Neil Goldsman, Chair/Advisor
Professor Pamela Abshire
Professor Thomas Antonsen
Assistant Professor Kevin Daniels
Professor Patrick McCluskey, Dean's Representative

Acknowledgements

I want to express my sincere gratitude to the people whose generous help and effort made my study in Maryland possible.

This work was supported by the Air Force Office of Scientific Research (AFOSR) Center of Excellence (COE) under Grant FA9550-15-1-0171.

I would like to thank Professor Thomas Antonsen for providing the funding opportunity for my research program. He offered both overall guidance and specific suggestions at all times, which kept my work moving forward and in good shape. Over the years, I received many chances from him to attend academic conferences and present my work, where I could learn from him and peer graduate students and researchers. Dr. Antonsen's engagement and input led me to see a bigger picture where my contribution can potentially benefit people working in various fields.

I would like to thank Professor Neil Goldsman, my research advisor, for mentoring me as I grow professionally. Dr. Goldsman does not hesitate to share his wisdom, energy, and inspiration ever since I became his student. I learned fundamental concepts and methodologies in semiconductor physics and real-world situations from him, many of which were beyond my graduate-level textbooks. He gave me chances to learn new things from time to time when I built up hands-on skills and gathered ideas to proceed with my study. He let me talk in front of our group and many other peer researchers, where I had the chance to present myself and practice my speaking skills. His incredible dedication to teaching influenced the way I communicate. When I was a teaching assistant for a few of his undergraduate classes, I learned to keep my content organized and motivating, making it easier for the audience to absorb. I perfected my circuit analytical and diagnostic skills when covering his electronic lab classes, which became vital when I started doing experiments for my thesis. He gave me rides in his electric car, which had me start thinking about what (a part of) electrical engineering and the semiconductor industry is all about. Although I had tense moments when he challenged my ideas and results, I appreciate his constant patience and attention as I corrected and improved myself. As I told him before:

He is a great mentor. His guidance is a beacon in the darkness as I explore the sea of knowledge. Being able to follow him over the past years has been a privilege and true pleasure. All my best wishes to him.

Besides my advisor and Professor Antonsen, I would like to thank the rest of my thesis committee: Professor Pamela Abshire, Professor Kevin Daniels, and Professor Patrick McCluskey, who all agreed to participate in my defense meeting on very short notice.

I would like to thank my peer students in Dr. Goldsman's group, many of whom have graduated: Chris Darmody, Christian Xiao, Franklin Nouketcha, Alex Mazzoni, and Ittai Baum. We developed the genetic algorithms upon each other's, which turned out powerful and handy. More generally, exchanging topics routinely gave me a broader and more profound vision of semiconductor physics, as well as the chance to think critically. Besides, our friendship outside the classroom made my life a joy.

Special thanks to Chris, who proofread my thesis and gave me professional input. He became my loyal audience and allowed me to bug him with uncommon questions about English grammar and idioms.

I would like to thank Dr. Akin Akturk for providing insightful comments which helped me understand and model the parasitic current mechanism.

Special thanks to Emily Irwin at our department. She promptly walked me through my graduation process, making the last part of my program much less stressful.

Finally, I would like to thank my mom and dad for their endless and unconditional parental love. Without their spiritual support and encouragement, it would not have been possible for me to live through all my years studying abroad without worries. Being their son is bliss. I hope they stay healthy and happy, and I hope we can spend more time together.

I met many people in the US who indirectly helped me with my study: Jingya, the Gao family, and Jiakun and Bin. Their companionship and generosity made my life delightful and meaningful. I truly cherish the moments we shared over the years.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Overview

Electromagnetic interference (EMI) commonly exists in electronic devices and systems containing semiconductor-based integrated circuits (IC's), or "chips". Disturbances in on-chip signals may affect their overall functionality and lead to device failure [1–3]. As IC's are more and more widely used in modern life, such as personal computers, mobile telecommunication devices, radio-frequency identification (RFID) and near-field communication (NFC) [4, 5], power supplies [6], automobile electronics [7, 8] and electronic medical devices [9, 10], their capability to sustain functionality under EMI is of continuous interest.

In this work, the vulnerabilities of silicon (Si) MOSFET devices ("devices" for short) to EMI are studied at the device and the circuit levels. When the printed circuit board (PCB) containing one or more IC's is experiencing EMI, the interference may couple into the tracings (metal wire routes) on the board [1]. The dimensions of the devices and supporting features in modern IC's ($\lesssim 1\,\mu\text{m}$) are typically much smaller than the interference's wavelength ($\approx 10\,\text{cm}$ for a 3 GHz wave in vacuum); thus, the EM waves possibly existing in or near the circuit usually do not directly interact with the devices.

However, the coupled EM interference may end up as localized transient and even resonant voltage and current disruptions at the terminal connections (pins) of the chip [11]. These disruptions may eventually reach the device and trigger non-linear responses due to parasitic behaviors other than the operational characteristics that MOSFETs are designed for.

As a result, the vulnerabilities induced by EMI may become out of proportion and exceed the designed tolerances such as the allowed minimum signal-to-noise ratio or max-

imum bit error rate. The circuit may even become trapped into an abnormal state giving erroneous outputs, e.g. digital signal upsets and bit-flips leading to unwanted latching [12]. These can lead to catastrophic events such as system hang-ups or crashes, requiring a cold reboot to reinstate the circuit's normal function. Additionally, under extreme circumstances, the EMI-induced disruptions may put up excessive stress on the devices — sometimes transient and unexpected by design, but the transient events can lead to permanent damage which the devices cannot recover from [13, 14].

We define two categories of EMI-induced vulnerabilities in MOSFET circuits according to the outcomes described above. The "Soft Errors" (SE) are non-permanent disruptions that can temporarily interrupt the functioning circuits or systems, such as generating bit flips in digital circuits or gain variations in analog amplifiers. While these events are detrimental, the circuits can be restored from the errors without long-term consequences. In Section 1.2, we introduce our simulation and experimental studies on one of the nonlinear behaviors in the devices, the Snapback phenomenon, which can lead to or intensify the Soft Errors.

The other type of vulnerabilities is the "Hard Failures" (HF) or permanent damages to the devices. Depending on the level of the EMI-induced voltage-current disruptions, a catastrophic event can occur in a very short time, comparable to the time scale of the interference, and the device will be permanently and severely damaged, irreversibly disabling the circuit's functions [15]; or there can be a long-term shift (degradation) in the device's performance and the circuit's viability, and the transient EMI disruptions can accelerate the process [16]. In Section 1.3, we introduce our work on the gate dielectric (oxide layer) breakdown, which is one type of Hard Failure, inculding experimental tests and simulations with nano-scale FinFETs.

Throughout the study of each of the two aspects, the vulnerabilities are modeled based on physical knowledge — as well as experiments that we performed — for silicon MOSFET devices. Device-level (or "Technology Computer-Aided Design" or "TCAD")

simulations are performed to further assist in the validation and modeling process. A device-level simulation solves distributed (position-dependent), partial differential equation-based systems of equations (as will be explained with more details in Equations 2.25), whereas a circuit-level simulation uses lumped components (only defining terminal characteristics) and algebraic (closed-form) model expressions. Two compact circuit models are developed in this work — one is for characterizing the Soft Errors with the Snapback phenomenon, and the other for the Hard Failures caused by oxide breakdown. Example applications of these circuit models are also demonstrated. The key parts in the proposed methodology are highlighted below:

• Experiments on individual MOSFET devices provide real-world observations of the vulnerability behaviors of our interest. The data collected from I-V measurements and oxide stress (breakdown) tests is analyzed statistically and then used to calibrate our proposed models. Our experiments serve as an essential "first step" of developing the physics-based models, which has the potential capability of using TCAD simulation data instead, once the understandings of the device vulnerability behaviors are obtained.

• Device-level simulations can bring insights into the device's internal behaviors, which are usually hard to observe experimentally in practical circuits and systems. The simulation tool, once calibrated, can accurately adapt to and predict the behaviors of various device designs without actually fabricating them.

• From the physics-based model derivation, compact models are made to produce simpler relationships between device terminal inputs and responses. The large amount of information — and computation — about the device from TCAD simulations is reduced. Thus, the evaluation of both Soft Errors and Hard Failures can be integrated into simulations of practical circuits.

• A method to calibrate the proposed model, i.e., to extract the model parameters, is developed based on a heuristic search method known as the Genetic Algorithm. This ex-

traction technique does not rely on function derivatives that are used in traditional optimization methods such as gradient descent, and it almost always guarantee a global minimum, so it is suitable for fast model prototyping of semiconductior behaviors, including but not limited to vulnerabilities. Once the necessary parameters are extracted, simulation-based vulnerability evaluation can guide the circuit design to improve the circuit's reliability.

1.2 Electromagnetic Interference-Induced Non-Permanent Vulnerabilities in Silicon MOSFET Devices ("Soft Errors")

The first part of the work focuses on non-permanent damages, or "Soft Errors" (SE) caused by transient disturbance in planar silicon MOSFETs under EMI conditions. In typical MOSFET circuit designs, it is widely seen that the drain and/or source terminals are connected to external power pins (V_{DD} and/or V_{SS}). Hence, these MOSFET devices connected to the power lines may experience different levels of EMI-induced voltage disruptions, depending on the actual circuit topology. This is often referred to as power integrity problems, and the level of terminal voltage disruptions induced by injected EMI is widely studied at the circuit and board levels using microwave-related techniques [1, 11]. It is easily understandable that when the power rail of a circuit is disrupted, there may be an increased amount of coherent noise leaking into the circuit's input and/or output signals. As mentioned before, possible consequences include undesired bit flips in digital circuits and deviations in analog circuit behaviors, which can propagate and lead to wider and more catastrophic problems such as system crash.

Circuit- and system-level studies on EMI-induced disruptions are widely reported in the literature. For example, in an experimental study by Mattei, et. al. [9], commercial implantable pacemakers were exposed to an EMI environment which simulated worst-case scenarios when RFID and NFC transceivers were operating close to emulated human bodies with implanted pacemakers. The RFID/NFC operation and dummy human body model followed established standards. The output of the system, the artificially generated heart pulses observed as voltage waveforms, was monitored. From the reported results, the tested pacemakers' performance was interfered with by the EM radiation used for *normal* RFID operations. The performance degradation disappeared after the RFID communication ended. A study by Sparks, et. al. [17] found similar temporary (transient) inhibition effects from GSM-band interference on implanted pacemakers. The EMI was supposedly

emitted from mobile telephones and base stations and enough to cause harmful disruptions, although the telecommunication industry generally receives much stricter governmental regulations, and cellphones generally emit less average power than common RFID devices.

In digital circuits, two of the most prominent Soft Error vulnerabilities under EMI are temporary bit errors due to input-output characteristic distortions [18, 19] and excessive power consumption due to additional leakage current [20]. As reported by Yang et. al. [18], the output bit error rate of a CMOS inverter may have a strong correlation with the level (amplitude), phase, and frequency of the injected EMI. In the experimental and simulation works by Kim et. al. [19], transient bit flips were able to propagate through cascaded logic gates (inverters) and result in bit errors. The less severe upsets — which may not cause propagation of bit errors — can still raise the noise level and influence the circuit's performance, such as clock jitters in synchronized logic circuits.

On the other hand, in battery-driven mobile digital communication devices and other portable systems, where the total average and transient power consumption is of particular interest, EMI-induced voltage-current disruptions may cause the MOSFET devices more conductive and dissipate more Joule power. In a study by Abedi et. al. [20], significant increases in measured leakage current (100x – 1000x) in CMOS inverters were reported when the injected EMI power increases by 12–15 dB (peak power $\sim$ 0 dB m or 1 mW). Non-linear increases in the power consumption were observed as the EMI level increased, which could be attributed to the MOSFET's structures and vulnerable behaviors.

In this work, we develop a methodology to evaluate the MOSFET circuit vulnerabilities triggered by EMI due to common intrinsic flaws in MOSFET devices that are unrelated to the main functions and are normally harmless. We study the internal, vulnerable structures in MOSFETs from various aspects, including physical mechanisms, numerical simulations, and experimental tests. We derive and calibrate a model to describe the disrupted input-output characteristics (steady state or DC) and transient responses of MOSFET devices in addition to regular behaviors. At the circuit level, our model represents the

MOSFET's internal vulnerabilities that can be activated by temporary EMI-induced disruptions. Thus, the model serves as a bridge between the EMI injection in the target circuits or "stimulus", and the MOSFET circuit's non-linear vulnerable behaviors or "response".

Unique to MOSFETs and similar transistor structures, several highly non-linear mechanisms collectively appearing as the "Snapback" phenomenon [21] may be activated by the disruptions in terminal voltages, including impact ionization [22]. Although under normal, designed conditions, these abnormal mechanisms do not change the device behaviors drastically, under disrupted conditions they can lead to or intensify Soft Errors in MOSFET circuits.

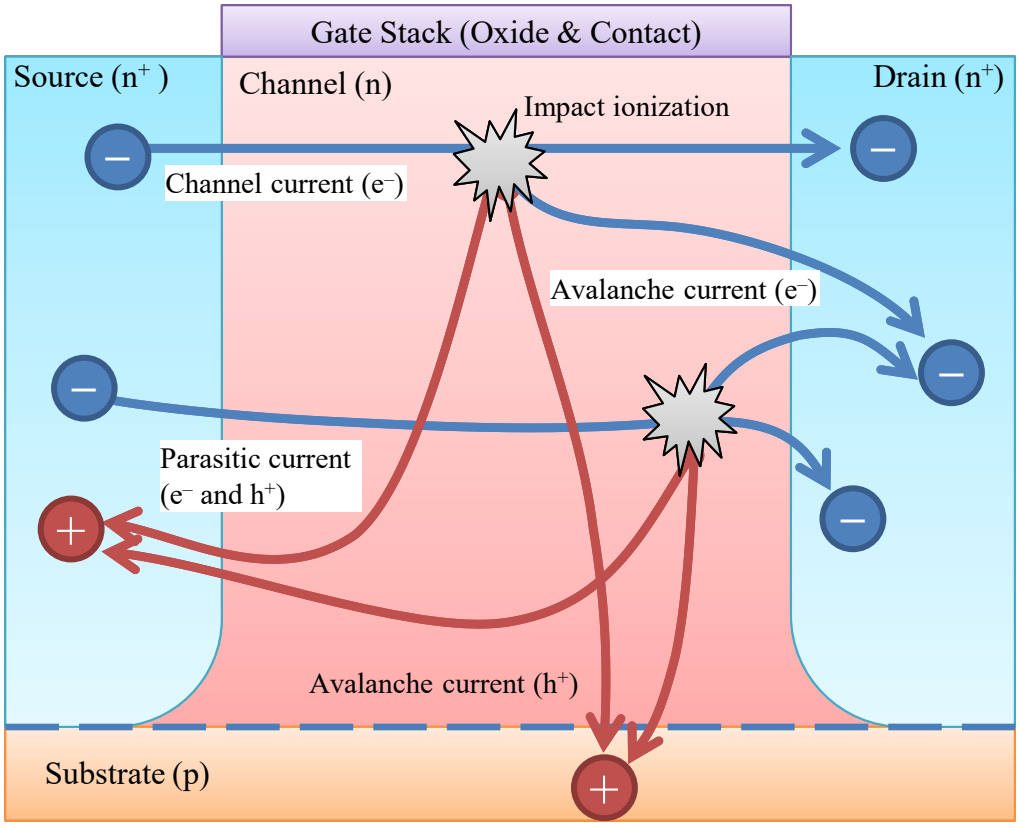


Figure 1.1: Illustration of current types in MOSFET used in the circuit model, considering impact ionization and parasitic structures. The formal description is in Section 2.3.3.

Figure 1.1 is a simplified illustration of the inner structure of an N-MOSFET and related vulnerability behaviors, which will be further investigated in Chapter 2. First, under high transient terminal voltages induced by external EMI [23], the channel current in an N-MOSFET increases rapidly (nearly exponentially) due to impact ionization, increasing the channel electron current and creating hole current in the substrate. Next, due to the existence of a parasitic structure similar to a bipolar junction transistor (BJT) inside the N-MOSFET, which is activated by the impact ionization-generated substrate hole current, the drain voltage decreases while the device experiences extremely high drain current. This phenomenon is known as *Snapback*, named after the "turning-back" pattern in the device current-voltage relation. When Snapback occurs, the device characteristics have non-monotonic solutions, and the related functional circuits are disrupted [24]. In extreme cases, the device can fall trapped in the low-channel-resistance state, uncontrolled by the gate (input) voltage, and hence its signal-level behavior may no longer follow design expectations. The consequences, not considering the self-heating due to excessive channel current, are mainly temporary upsets, such as signal distortions or bit flips, which are in addition to the erroneous results calculatable with regular MOSFET models. In general, Snapback is the extreme situation worsening the signal integrity. In digital logic gate circuits, for example, temporary upsets may occur at the signal or software level, potentially leading to a calculation error or system crash. As a consequence, the circuit may need to be cold-rebooted (or reset) to return to normal function.

Modeling impact ionization and Snapback, as well as the related electrostatic discharge (ESD) vulnerability and latch-up in CMOS devices has attracted a lot of attention in research and practical applications [21, 25–29]. Equivalent circuits containing parasitic bipolar current and avalanche current using empirically determined exponential I-V relationships were proposed [21, 25–27]. Commonly used compact models contained avalanche current expressions that were heavily dependent on empirical data and calibration, which could largely vary due to fabrication process differences such as gate length and

doping concentration, although geometry scaling was included in some proposed models [26, 27]. A testing standard was proposed to trigger and detect any possible signal latch-ups in digital CMOS circuits [28] with a pulsed ($\sim 10\,\text{ms}$) current of $100\,\text{mA}$ or a pulsed voltage of $50\,\%$ over the operating power rail. Generally speaking, impact ionization and Snapback will cause more severe problems in smaller devices, since the electric field is generally higher, leading to larger electron-hole generation rates. One feasible way to accurately capture the detailed behaviors of these vulnerability-related effects is to perform experimental tests with fabricated devices and extract the SPICE model parameters [29]. Meanwhile, device and circuit-level simulations using calibrated models can largely reduce the cost of designing future prototypes.

In this study, a workflow using multi-level simulation based on physical models is developed. Experimental tests are conducted to verify and calibrate the simulation model. First, we experimentally measure the Snapback phenomenon in standalone MOSFETs. We also perform device-level simulations to validate the physical mechanisms and their quantitative effects. Next, using physics- and geometry-based expressions, we derive a compact circuit-level Soft Error model. By using our in-house developed techniques, the compact model parameters are extracted from our experimental data. In addition, we discuss possible methodology improvements in our detailed device-level simulations, in an attempt to substitute future experiments with computational data for the model extraction. The compact model is compatibile with regular circuit simulators, so it can evaluate a circuit's and system's vulnerability — or reliability — and provide design improvements. Thus, we simulate practical MOSFET circuits under EMI-induced voltage-current disruptions using our calibrated Soft Error model. Through post-simulation analyses, we quantitatively relate the level of vulnerability to the level of injected EMI.

In Chapter 2, a physics-based model of impact ionization (the lucky-drift model [30]) is investigated. The lucky-drift model analytically describes local generation due to impact ionization. A silicon N-MOSFET using a planar $0.18\,\mu\text{m}$ process is simulated at the

device level. Impact ionization-induced local generation of electron-hole pairs is observed, which contributes to increased drain current. The generated holes in the body/substrate region lead to locally elevated potential near the body-source junction. The parasitic BJT structure, which exists in a regular MOSFET, is thus activated and can significantly increase the drain current in addition to the impact ionization. Together, these phenomena can finally lead to Snapback, when the uncontrollable secondary current dominates the drain-source conduction, and the device falls into a low-resistance state.

We create a reduced circuit-level model for impact ionization and Snapback for a single N-MOSFET device [26] using compact or closed-form equations that we derived by summarizing the device-level phenomena. All necessary parameters in this model can be extracted from terminal characteristics, given the knowledge of the device's physical structure. Together with a regular MOSFET model, a compact Soft Error model is established, and a Snapback event is recreated in our simulation.

In Chapter 3, we calibrate the proposed model with experimental data. Snapback experiments have been carried out, and we have measured the impact ionization and Snapback currents in commercially available N-MOSFET devices. The Soft Error model parameters are extracted with our in-house methodologies based on a Genetic Algorithm.

Finally, using the extracted compact Soft Error model, we can perform circuit-level simulations with N-MOSFET-based circuits. In Chapter 4, a potential application of the proposed Soft Error model, based on the Snapback phenomenon, is demonstrated with more details. Under the influence of transient EMI disruptions, we observe and analyze the changes in circuit behaviors.

1.3 Electromagnetic Interference-Induced Permanent Vulnerabilities in Silicon MOSFET Devices ("Hard Failures")

The second part of this work studies the permanent damages, or "Hard Failures" (HF) in silicon devices. Similar to the previously discussed Soft Errors, MOSFET circuits under EMI conditions can experience transient voltage-current disruptions. These temporary upsets can stress the devices extensively and cause permanent damages, and the fundamental device behaviors may change even after the stress conditions are removed and the system is rebooted, unlike the temporary failures as in the Soft Error case.

Among a broad range of permanent damage mechanisms leading to Hard Failures, gate dielectric (or "oxide" as in SiO_2) breakdown (BD) may occur under high gate terminal voltages (V_{GS}, V_{GD}, or V_{GB}). Modern MOSFET devices generally have oxide layer thickness of 2–25 nm for CMOS digital logics, and up to 50 nm for power devices. Normally, the gate oxide behaves as the insulating dielectric portion of a capacitor so that the charge stored near the semiconductor substrate interface can be used for current conduction. Thanks to the oxide's insulating nature, MOSFET-based circuits have advantages over BJT-based ones such as higher input impedance, higher fan-out, and lower quiescent power dissipation. When the oxide breaks down, significant leakage current appears across the oxide, which becomes resistive; effectively, shunt resistors come to exist between the gate and other terminals (source, drain, and body), leading to complications such as additional parallel resistive load (even undesired short-circuits) and thermal runaway [15]. The changes in circuit behaviors, including catastrophic failures, can permanently disable the circuit's functions. The increased gate leakage current, together with possible increased channel current (from drain-source stress-induced impact ionization or parasitic current), can lead to extreme Joule heating, damaging the device's lattice structure.

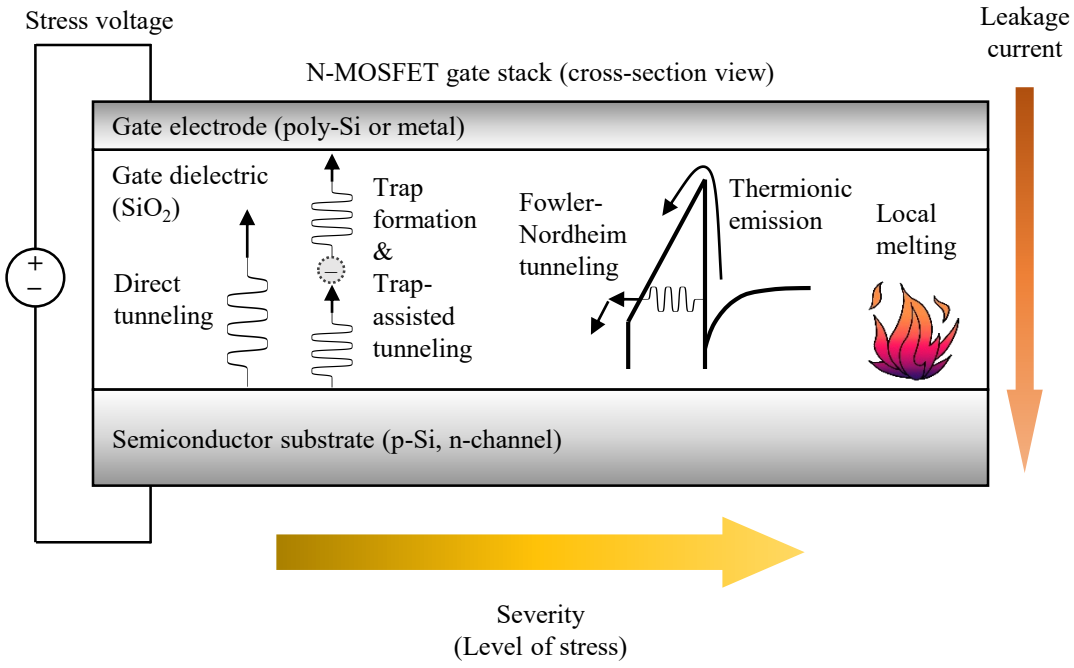


Figure 1.2: A conceptual illustration of oxide breakdown mechanisms. Several mechanisms are listed, as the severeness intensifies and the stress condition worsens. The rectangular boxes show a typical planar MOSFET's gate stack structure (substrate-dielectric-electrode). The simplified dielectric-substrate band structure is for illustrational purposes only and not aligned with the gate stack drawings. The arrows indicate the flow of electrons when stress voltage and leakage current are present, and the electric current flows from the gate electrode (top) towards the semiconductor substrate (bottom).

Figure 1.2 is a conceptual illustration of several possible damaging mechanisms to the gate dielectric, shown on top of a simplified N-MOSFET gate stack. The finite potential barrier formed by the oxide-semiconductor interface allows for quantum tunneling current, giving rise to the leakage current that is ubiquitous in all devices under applied voltage. The direct tunneling current is the most common type as it does not require additional mechanisms other than the channel electron wavefunction penetrating into the dielectric's bandgap. The tunneling current magnitude can be evaluated with analytical methods (e.g., the Tsu-Esaki model and the WKB approximation [31, 32]) and numerical methods (such as the transfer matrix method [33, 34] and the quasi-bound state method [35–37]). The simulated results for gate tunneling current density are around $J_G =$

1–10 $\mathrm{nA\,\mu m^{-2}}$ (or 0.1–1 $\mathrm{A\,cm^{-2}}$) for dielectric thickness of 1–2 nm at around 1–2 V at room temperature, neglecting the self-heating effect which is highly likely to cause melting at this level. In fact, from the heat equation and ignoring the heat flux, one can estimate the time to reach the oxide melting temperature for a thin oxide ($t_{OX} = 1\,\mathrm{nm}$) due to local heating by $t_{\mathrm{melt}} = \frac{T_{\mathrm{melt}} - T_{\mathrm{initial}}}{W/C\rho} = 0.45\,\mathrm{ms} \times \frac{1\,\mathrm{A\,cm^{-2}}}{J_G}$ where $C = 1 \times 10^3\,\mathrm{J\,kg^{-1}\,K^{-1}}$, $\rho = 2.65 \times 10^3\,\mathrm{kg\,m^{-3}}$ are the specific heat and mass density of SiO_2, $T_{\mathrm{melt}} = 1710\,^{\circ}\mathrm{C}$ and $T_{\mathrm{initial}} = 30\,^{\circ}\mathrm{C}$ are estimated melting and normal temperatures, and $W = 1\,\mathrm{Volt} \times \left(\frac{J_G}{t_{OX}}\right) = 1 \times 10^{13}\,\mathrm{W\,m^{-3}} \times \left(J_G \div 1\,\mathrm{A\,cm^{-2}}\right)$ is a rough estimation for heating power density. This is concerning for ultra-thin gate dielectrics.

Under a high level of gate voltage stress, the potential barrier becomes generally thinner, and Fowler-Nordheim tunneling may occur [32, 38], when hot channel electrons (or holes) with kinetic energy much higher than the average thermal energy can tunnel into the dielectric's conduction (or valance) band, adding to the leakage current. This type of tunneling is generally smaller than the direct type under normal operation conditions or moderate stress, since it requires higher-energy carriers, which are less popular. However, as the gate and/or drain stress increases, it may increase faster and exceed the direct tunneling in thicker dielectrics when a high applied voltage causes the trapezoidal potential barrier become a triangular one.

Under high drain voltage stress, channel carriers may obtain enough kinetic energy to directly cross the potential barrier, forming thermionic emission current, as can be observed from a simulation study [39] in agreement with experimental data. In the cited work, by solving the energy balance equation for the non-equilibrium energy distribution in the channel, the leakage current is calculated by counting the "hot" electrons with enough kinetic energy to surpass the potential barrier.

In all cases, the elevated amount of gate tunneling (or leakage) current can cause local heating effects in or near the gate dielectric layer, which can be worsened when the channel (drain-source) conduction current is multiplied by the avalanche effect due to im-

pact ionization under high channel field. As a consequence, local melting of the dielectric layer or the underlying semiconductor substrate may happen [14], leading to catastrophic, permanent device destruction.

On the other hand, the electrons tunneling into the dielectric's bandgap (the forbidden band) can become trapped by local intrinsic defects. The intrinsic defects are local sites with irregular atom arrangements that naturally exist in thermally grown dielectrics, such as the amorphous SiO_2 network. Through simulation studies, typical O–Si–O bonds were found to have angles of around 109° [40] while the wide bond angles ($> 130°$) also randomly exist in the network with a concentration of $4 \times 10^{19}\,\mathrm{cm}^{-3}$ [41], which translates to $\sim 1 \times 10^{13}\,\mathrm{cm}^{-2}$ in a 2 nm thick layer. These defects can provide additional local states with energy levels in the oxide bandgap and near or below the Si conduction band bottom. They are capable of "capturing" channel electrons that tunnel into the SiO_2 via direct tunneling or phonon-assisted tunneling [42].

There are at least three consequences related to this local trapping of electrons. First, the trapped electrons negatively charge the gate dielectric, changing the device's threshold voltage (ΔV_{TH}) and its electrical behaviors [43]. Depending on the stress conditions, this change may be temporary and recoverable — when the stress alternates between positive and negative voltages — or permanent and able to accumulate until the device fails to operate if a unipolar stress continuously adds charge to the dielectric. Second, the local defect state in the dielectric's bandgap can serve as an intermediate stop for the electron tunneling from the channel towards the gate electrode by a capture-emission process, promoting more tunneling current categorized as the trap-assisted tunneling (TAT) current[1] [42]. Third, the extra electrons captured by the local defects can adversely affect the defects by catalyzing a chemical process creating oxygen vacancies [44]. It is believed that

[1] It is worth noting that, in the cited work [42], the experimental data used for model calibration seemingly reported that their SiO_2 layers operated under fields of over $10\,\mathrm{MV/cm}$ before a catastrophic breakdown, which is higher than usually observed. It might be because the formula used for the field calculation overestimated or because the dopants, undocumented, affected the threshold voltage in the stress-tested devices. After all, the simulation result may have uniformly underestimated the leakage current, although the proposed methodology has brought insights to the leakage mechanisms.

these stress-induced oxygen-vacancy defects can promote more TAT leakage current due to different defect energy levels, while self-heating can accelerate the reaction. Thus, the electron-assisted growth of tunneling current can be a major contributor to the total leakage current. These defects are permanent and can increase the tunneling current even after the high-level stress, leading to the defect creation, is removed.

Tunneling current of any kind can eventually become significant, making the gate dielectric layer effectively conductive, thus rendering the MOSFET in a degraded or even non-functioning state. The traps created under high stress remain in-place after the stress is removed, thus causing a "memory" effect of higher leakage current, which can become enough to cause catastrophic damages such as local melting, if such destructive result has not occurred under extremely high-level of stress.

In Chapter 5, we relate the gate dielectric BD with measurable, circuit-related quantities, such as applied gate voltage and DC conduction current, in an attempt to predict the BD event by evaluating terminal conditions during a circuit simulation. First, we review the BD mechanism of SiO_2 in Section 5.1. One of the critical factors determining the onset of BD events is the electric (E) field in the oxide [45]. A threshold value for instant or rapid, permanent BD in SiO_2 dielectrics is typically 8–10 $\mathrm{MV/cm}$.

We fabricated test devices at the NanoCenter FabLab, University of Maryland, and experimentally stressed them to extract BD conditions such as the threshold E field. We describe our fabrication process and experiments in Sections 5.2, 5.3 and 5.4. Base on the experimental data, we establish the criteria for rapid oxide BD, which will be used later for the Hard Failure circuit-level simulation model.

Meanwhile, as the device dimensions keep scaling down as a general trend, the gate oxide becomes thinner and generally faces more difficulties in field-driven BD. For example, the constant-field or constant-power-density scaling of CMOS digital circuits [46] dictates that the device width, length, effective oxide thickness, and power rail voltage scale down by the same factor so that the channel (lateral) field, the gate dielectric (perpendicu-

lar) field, and the power per device area are maintained. The fabrication processes at present typically have gate lengths of $L_G = 10$–$25\,\text{nm}$ and oxide thicknesses of $t_{OX} = 2$–$10\,\text{nm}$ or up to 50 nm for power devices.

However, the disruption voltages induced by external EMI can still be of the same level. Proportionally, it may be easier for oxide BD to happen (e.g., a 1 nm-thick SiO_2 layer may break at around 1 V or less). In Table 1.1, we summarize selected key design factors from the literature on the CMOS digital circuit scaling trend. We also give rough estimations of possible peak electric field in the gate dielectrics (assuming HfO_2) under normal working conditions (E_{OX}) and disrupted conditions (E_{1V}) of 1 V, possibly induced by external EMI. The estimated physical gate oxide (HfO_2) layer thickness is generally $t_{OX} = 2$–$3\,\text{nm}$, and when it is operating under normal conditions, the peak oxide electric field is $E_{OX} = 2$–$3\,\text{MV/cm}$. When it is under 1 V of disrupted voltage, which can be induced by EMI, the electrical field can be as high as $E_{1V} = 4\,\text{MV/cm}$. The BD field for HfO_2 is similar to SiO_2 [47, 48] (to be discussed in Section 5.1). Thus, EMI-induced disruptions and stress conditions may still be a vulnerability concern.

Table 1.1: Selected key device scaling factors reported from literature (high-performance logic or "HP") and estimated peak electric field in the gate dielectric layer

Production year	2010		2015-2016			2019		2020-2021			2024	2025-2027		2030-2031	
Technology node ["nm"]	"45"		"22"		"16/14"	"7"		"6/5"	"5"		"3"	"2.1"		"1.5"	"1.0"
Selected key parameters in the referenced studies															
Reference	[49]	[50]	[49]	[50]	[51]	[51]	[52]	[51]	[52]	[53]	[52]	[52]	[53]	[52]	[53]
L_G [nm]	18	30	9	22	24	14	18	10	16	18	14	12	14	12	12
W_{eff} [nm]					92	56.5	107	56.5	72	126	96	110	192	24	160
C'_{TOT} [fF/µm]	0.57		0.35		1.81	1.29	1.27	0.97	1.42	0.45	1.29	0.95	0.39	0.89	0.37
V_{DD} [V]	1.0		0.80			0.70		0.65		0.70	0.65			0.60	
Estimated gate effective oxide thickness using SiO_2 formula (* From reference)															
EOT [nm]	1.1	1.5*	0.9	1.2*	0.5	0.4	0.5	0.4	0.4	1.4	0.4	0.4	1.2	0.5	1.1
Estimated gate oxide thickness and peak gate oxide field under normal operation conditions using HfO_2 formula (* From reference)															
T_{OX} [nm]	7.0	4*	5.7	2.5*	2.9	2.4	3.1	2.3	2.5	8.9	2.4	2.8	7.9	3.0	7.2
E_{OX} [MV/cm]	1.4	2.5	1.4	3.2	2.7	2.9	2.2	2.8	2.6	0.8	2.7	2.3	0.8	2.0	0.8
$E_{1\text{V}}$ [MV/cm]	1.4	2.5	1.8	4.0	3.4	4.2	3.2	4.4	4.0	1.1	4.2	3.6	1.3	3.4	1.4

Symbol	Meaning	Unit
L_G	Physical gate length	nm
W_{eff}	Effective channel width	nm
C'_T	Gate capacitance per width	fF/µm
V_{DD}	Power rail voltage	V
EOT	Estimated equivalent SiO_2 thickness	nm
T_{OX}	Estimated equivalent HfO_2 thickness	nm
E_{OX}	Estimated operation gate field	MV/cm
$E_{1\text{V}}$	Estimated disrupted gate field	MV/cm

Estimation Formulas

$$\text{EOT} = \varepsilon_{\text{SiO}_2}/C'_G$$
$$T_{\text{OX}} = \varepsilon_{\text{HfO}_2}/C'_G$$
$$E_{\text{OX}} = V_{DD}/T_{\text{OX}}$$
$$E_{1\text{V}} = 1\,\text{V}/T_{\text{OX}}$$
$$C'_G = C'_T/L_G$$
$$\varepsilon_{\text{SiO}_2} = 3.9\varepsilon_0$$
$$\varepsilon_{\text{HfO}_2} = 25\varepsilon_0$$
$$\varepsilon_0 = 8.85 \times 10^{-12}\,\text{F/m}$$

In Chapter 6, one example device structure, silicon FinFET, is studied for its vulnerabilities from the gate oxide BD perspective. The investigated structure utilizes an ultra-thin (3 nm) gate layer dielectric of SiO_2. Silicon FinFETs are one type of the newest-generation devices [54]. With their unique vertical channel structure, they can achieve better performance than planar devices of comparable channel length [55], giving them an important role in the race of device scaling [52]. A FinFET device (N-MOSFET) with basic functional structures is simulated at the device level to observe its electrical behaviors as a baseline. Quantum-mechanical effects are considered by solving the time-independent Schrödinger's equation and the electrostatic Poisson's equation self-consistently, giving a corrected result for the gate oxide field. Next, we extend the knowledge from our SiO_2 BD experiments to the thinner and much smaller FinFET structure. A Hard Failure circuit model representing the rapid oxide BD is proposed for the simulated FinFET, combining our experiments and simulations.

Chapter 2: EMI-Induced Soft Error Vulnerabilities: Simulation Model of the Snapback Phenomenon

Soft Errors are temporary, non-permanent damages in circuits containing MOSFET devices. EMI-induced voltage and current disruptions may lead to highly non-linear behaviors in MOSFETs. One of the main related issues is the Snapback phenomenon, which will be studied in this chapter.

The Snapback phenomenon itself is related to several physical mechanisms, including impact ionization and the parasitic bipolar-junction (BJT) structure. A Soft Error simulation model (the "Soft Error model") is proposed based on the Snapback phenomenon. The model acts as an add-on module to a typical N-type MOSFET, and it is compatible with SPICE circuit simulations.

In Section 2.1, we use a probability model to describe impact ionization as a scattering mechanism, treating electrons as particles. The local ionization rate ("*avalanche rate*") describes the number of generated (due to impact ionization) carriers per unit length, which will be converted to the per-volume generation rate G_A, and is given by

$$\alpha(E) = \frac{qE}{\mathscr{E}_i} \exp\left(-\frac{\mathscr{E}_i}{q\lambda_R E}\right) \tag{2.1}$$

where E is the local electric field magnitude, $\mathscr{E}_i$ is the ionization threshold energy, λ_R is the average mean free path of phonon scattering, and q is the electron charge.

It can be shown [56] that the exponential term in Equation 2.1 represents the fraction of the total conduction-band electron population that has kinetic energy above the threshold energy $\mathscr{E}_i$. Comparatively, the fraction of the above-threshold ($\mathscr{E}_i$) electron population following the Maxwell-Boltzmann distribution at equilibrium at temperature T is given by a definite integral of the distribution function $f_{\mathscr{E}}(\mathscr{E})$ as

$$\int_{\mathscr{E}_i}^{+\infty} f_{\mathscr{E}}(\mathscr{E})\, d\mathscr{E} = \frac{2}{\sqrt{\pi}} \sqrt{\frac{\mathscr{E}_i}{kT}} \exp\left(-\frac{\mathscr{E}_i}{kT}\right) \tag{2.2}$$

Therefore, by comparing the exponential terms in Equations 2.1 and 2.2, under the acceleration force from external field E, the "variance" term kT is replaced by the average work done by the external field $q\lambda_R E$ (and the average energy $\frac{3}{2}kT$ is replaced by $\frac{3}{2}q\lambda_R E$), and effectively, the equivalent temperature of the 3-dimensional electron gas is changed to

$$T_e = T_e(E) = \left(\frac{q\lambda_R}{k}\right) E \tag{2.3}$$

For example, with an estimated mean free path of $\lambda_R = 20\,\text{nm}$ [57], under external field $E = 50\,\text{kV/cm}$, the equivalent electron temperature is $T_e = 1160\,\text{K}$, which is much higher than the lattice temperature (at about room temperature).

Impact ionization is important to us since it creates more carriers (and thus more current) under higher field. In other words, under disrupted high voltages, a MOSFET may experience excessive channel current from avalanching. Other parasitic effects may also be triggered, complicating and intensifying the situation. Our first step to modeling MOSFET vulnerabilities under EMI is to simulate impact ionization at the device level using the semiconductor's drift-diffusion equations.

To couple the impact ionization into the drift-diffusion equations (Equations 2.25), the avalanche rate is converted to a local generation term G_A (or number of generated carriers per unit volume per unit time) as

$$G_A = J\ \alpha(E) \tag{2.4}$$

where J is the local current density magnitude, and E is the electric field component in the same direction as the local current. We will further discuss the calculation in Section 2.1.

Next in Section 2.2, a single MOSFET is simulated at the device level. The effects of impact ionization are observed inside the device. The parasitic bipolar structure is discussed. Based on the results, vulnerabilities caused by the impact ionization and parasitic structure are reduced to compact circuit models.

We summarize the aforementioned impact ionization-related vulnerabilities with the Soft Error model, leading to an equivalent circuit of a single MOSFET, shown in Figure 2.1. In Section 2.3, we derive the circuit-level compact model for Soft Errors in N-MOSFET and verify it with device-level DC simulations.

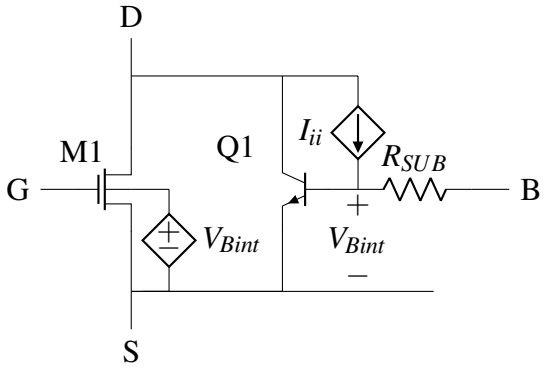


Figure 2.1: Full equivalent circuit for impact ionization and parasitic bipolar behaviors. All components are available using compact models in a circuit-level simulator. The detailed description is in Section 2.3.

In Chapter 3, the model parameters are extracted from a realistic N-MOSFET device, and Soft Error vulnerabilities in practical circuits are further investigated. In Section 3.1, we have experimentally measured the DC characteristics of a commercially available N-MOSFET device. In Section 3.2, we simulate the tested N-MOSFET at the device level, after recreating its structure based on extracted parameters.

In Section 3.3, we provide an extraction method for model parameters using a Genetic Algorithm, and we extract a Soft Error model for our experimentally tested devices. Finally in Section 4, we apply the Soft Error model in functional circuits and simulate their transient behaviors under EMI. Bit errors and possible system latch-up conditions are demonstrated.

By using this methodology, a scalable device model including Snapback is created and then reduced to a few circuit components. The model can be calibrated with experimental measurements or device-level simulations, while the parameters may reflect the physical aspects of various designs.

2.1 Physics-Based Modeling of Impact Ionization

The impact ionization process in silicon MOSFET channels is associated with scattering events involving high-energy electrons (hot carriers) and creating electron-hole pairs. We describe the impact ionization with a universal geometric probability model and the phonon scattering events in electron transport [57]. In this section, we use the symbol $\mathscr{P}(\cdot)$ for unitless, normalized probabilities.

First, we start with the basic concepts of scattering events in electron transport. The average scattering rate per spatial distance can be used as geometric probabilities. Assume the mean free path of an arbitrary scattering type, or the average distance between two "collisions", is an arbitrary value λ. The probability of an electron *not scattering* in an infinitesimal region Δx as it travels through is

$$\mathscr{P}(\text{no scatter in region } [0,\Delta x]) = (1-\Delta x/\lambda) \tag{2.5}$$

Within a region of $[0,x]$ consisting of a total number of N consecutive small segments, the probability for the electron *not scattering* as it travels through is equal to the total probability of it not scattering in any of the segments.

$$\mathscr{P}(\text{no scatter in region } [0,x]) = \prod_{k=1}^{N}(1-\Delta x_k/\lambda_k) \tag{2.6}$$

Assuming all small segments are identical and independent spaces, all λ_k become identical, all Δx_k become identical, and $N = x/\Delta x$. Therefore, Equation 2.6 becomes

$$\mathscr{P}(\text{no scatter in region } [0,x]) = \prod_{k=1}^{N}(1-\Delta x/\lambda) \tag{2.7}$$

Taking the logarithm on both sides of Equation 2.7, we have

$$\begin{aligned}\ln \mathscr{P}(\text{no scatter in region } [0,x]) &= \ln \prod_{k=1}^{N}(1-\Delta x/\lambda) \\ &= \sum_{k=1}^{N} \ln\left(1-\frac{\Delta x}{\lambda}\right) \\ &\approx \int_0^x -du/\lambda = -\frac{x}{\lambda}\end{aligned} \tag{2.8}$$

Therefore, Equation 2.7 becomes

$$\mathscr{P}\left(\text{no scatter in region } [0,x]\right) = \exp\left(-x/\lambda\right) \tag{2.9}$$

The probability density function $f(x)$ is defined as the (per-unit-length) probability of traveling for a non-infinitesimal distance x without scattering, and scattering at the end. This quantity is necessary for the next steps, since impact ionization requires a non-infinitesimal distance of uninterrupted acceleration so that the electron can gain enough kinetic energy from the external electric field to initialize the ionization event.

$$\begin{aligned} f(x)\,dx &\overset{\text{def}}{=} \mathscr{P}\left(\text{no scatter in } [0,x] \text{ and scatter in } [x,x+dx]\right) \\ &= \mathscr{P}\left(\text{no scatter in } [0,x]\right)\ \mathscr{P}\left(\text{scatter in } [x,x+dx]\right) \\ &= \left[\exp\left(-\frac{x}{\lambda}\right)\right]\frac{dx}{\lambda} \\ &= \frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right)dx \end{aligned} \tag{2.10}$$

The associated cumulative density function $F(x)$ describes the (unitless) probability of having a scattering event *somewhere* within a non-trivial distance from 0 to x as

$$F(x) = \mathscr{P}\left(\text{scatter in } [0,x]\right) = \int_0^x f(x)\,dx = 1 - \exp\left(-\frac{x}{\lambda}\right) \tag{2.11}$$

Now we can apply the above "generic" concepts to the impact ionization process. We may consider impact ionization as a scattering mechanism with a mean free path λ_i, although it is different than phonon scattering events because the former refers to the interaction between a high-energy mobile carrier (conduction-band electron or valence-band hole) and a valence-band electron (not conducting since it is in the inner shell), while the latter is the energy (and momentum) exchange via phonons between the electrons and the crystal lattice. Nevertheless, they are interrupting events in the electron transport process, and energy transfer is involved.

Directly determining λ_i is possible through simulations and experiments [58–61]. On the other hand, it is also possible to represent λ_i in terms of the mean free path of the phonon scattering events, which are typically much more common. The *lucky-drift model* [30] is one example of replacing λ_i with another term λ_R, or the average phonon scattering

mean free path[2]. Thus, the phonon scattering mechanisms of many possible types (and many other scattering mechanisms) [57] are summarized to one type with one mean free path value λ_R and one phonon energy value $\mathscr{E}_R$. Typically and for Si, $\lambda_R \ll \lambda_i$. We define a unitless quantity $r = \frac{\lambda_i}{\lambda_R}$, and $r \gg 1$.

The lucky-drift model in its simpliest form is based on two assumptions. While traveling, the carrier gains energy from the external field. Assume that under a low field, a carrier (presumably in the MOSFET channel) must start its transport process with a very small random energy $\mathscr{E}_0$, where $\mathscr{E}_0 < \mathscr{E}_R$ ("cold-electron" and low-field assumptions). We also assume that after a phonon scattering event, the carrier thermalizes and loses all its energy gained from the external field, and starts over with a new transport process (no-multi-stage-process assumption).

The impact ionization mechanism can be described in two steps. First, the carrier must reach the minimum energy required to initiate an impact ionization event, or the *ionization threshold energy* $\mathscr{E}_i$, which is approximately three-half times the band gap or $\mathscr{E}_i \approx \frac{3}{2}\mathscr{E}_g$, but could be different depending on the material's conductivity effective mass for electrons and holes. Because $\mathscr{E}_i \gg \mathscr{E}_R$, an average electron must avoid the phonon scattering to gain this energy while it is accelerating under the external field. Under homogeneous field E, the minimum uninterrupted distance for an electron reaching the threshold energy $\mathscr{E}_i$ is approximately $x = (\mathscr{E}_i - \mathscr{E}_0) \Big/ qE \approx \mathscr{E}_i \Big/ qE$. Therefore from Equation 2.11,

$$\begin{aligned} p_1 &\stackrel{\text{def}}{=} \mathscr{P}\,(\text{energy becomes above threshold}) \\ &= \mathscr{P}\,(\text{no scatter in } x) \\ &= 1 - F(x) \\ &= \exp\left(-\frac{\mathscr{E}_i}{q\lambda_R E}\right) \end{aligned} \tag{2.12}$$

Immediately after the electron energy becomes above the threshold, it either thermalizes through a phonon scattering event and consequently loses all its energy, or initiates

[2]The subscript "R" refers to Raman spectroscopy, which is an experimental method by which one can measure the phonon energy $\mathscr{E}_R$.

an impact ionization event. The probabilities of the two events can be associated to the scattering rates per infinitesimal length, or

$$R_i \stackrel{\text{def}}{=} \frac{1}{\lambda_i} \tag{2.13}$$

$$R_R \stackrel{\text{def}}{=} \frac{1}{\lambda_R} \tag{2.14}$$

Since R_R represents the scattering rate of all types except impact ionization, the total scattering rate is $R_{\text{total}} = R_i + R_R$. Hence, the *conditional probability* of having an impact ionization event after gaining the threshold energy is

$$\begin{aligned} p_2 &\stackrel{\text{def}}{=} \mathscr{P}(\text{ionize before phonon scattering}|\text{energy becomes above threshold}) \\ &= \frac{R_i}{R_{\text{total}}} = \frac{R_i}{R_R + R_i} \\ &= \frac{\lambda_i^{-1}}{\lambda_R^{-1} + \cancel{\lambda_i^{-1}}} \approx \frac{\lambda_i^{-1}}{\lambda_R^{-1}} \ (\lambda_i \gg \lambda_R) \\ &= \frac{\lambda_R}{\lambda_i} = \frac{1}{r} \end{aligned} \tag{2.15}$$

After all, the probability of an electron ionizing in a tentative distance of $x = \mathscr{E}_i \Big/ qE$ under external field is the product of p_1 in Equation 2.12 and p_2 in Equation 2.15.

$$p_3 \stackrel{\text{def}}{=} \mathscr{P}(\text{ionize in } x) = p_1 p_2 = \frac{1}{r} \exp\left(-\frac{\mathscr{E}_i}{q\lambda_R E}\right) \tag{2.16}$$

To apply this result to the Drift-Diffusion system of equations, we need to convert it to a *probability density* (per unit length), and so it can be used for the local generation terms.

Consider a cylindrical space of length $x = \mathscr{E}_i \Big/ qE$ in the MOSFET channel. The average time of flight for a carrier transporting through can be estimated with the average drift velocity v_E under field E, that is

$$t = \frac{x}{v_E} \tag{2.17}$$

Assuming the total volume of this cylinder is V, and the total mobile charge in it is Q, the *generation rate* G, or the average number of electron-hole pairs created due to impact ionization per unit time per unit volume, is given by

$$
\begin{aligned}
G &= \left(\frac{Qp_3}{q}\right)\frac{1}{tV} = \left(\frac{Q}{qV}\right)\frac{p_3}{t} \\
&= (n)\frac{p_3}{x/v_E} = (nv_E)\frac{p_3}{x} \\
&= \frac{J}{q}\frac{p_3}{x}
\end{aligned}
\tag{2.18}
$$

where $n = Q\Big/qV$ is the local mobile charge concentration, $J = qnv_E$ is the local current density magnitude, and q is the elementary charge. Therefore, we can define the "avalanche rate" or "ionization rate" $\alpha(E)$ as

$$
\alpha(E) \stackrel{\text{def}}{=} \frac{p_3}{x} = \left(\frac{1}{r}\right)\frac{qE}{\mathscr{E}_i}\exp\left(-\frac{\mathscr{E}_i}{q\lambda_R E}\right) \tag{2.19}
$$

One may apply different λ_R and even different $\mathscr{E}_i$ for electrons and holes[3], respectively, yielding α_n for electrons and α_p for holes. So the generation rate in Equation 2.18 can be rewritten for electrons (G_n) and holes (G_p) as

$$
\text{Local generation rate:} \qquad G_n = G_p = \frac{1}{q}\left|\vec{J}_n\right|\alpha_n(E_n) + \frac{1}{q}\left|\vec{J}_p\right|\alpha_p(E_p) \tag{2.20}
$$

Both electron-initiated (α_n) and hole-initiated (α_p) impact ionization are included. Note that there are equal number of electrons and holes generated by impact ionization, so $G_n = G_p$.

In the lucky-drift model, the "E-field" E and the "current density" $J_{n,p}$ ("n" for electrons and "p" for holes, respectively) are scalars because the formula is derived in a one dimensional space. To adapt it to 2D and 3D calculations, we need to extend the definition so the local generation rate in Equation 2.20 can be incorporated to the drift-diffusion system of equations (Equation 2.25). There is more than one way to redefine $\alpha_{n,p}(E)$ and $G_{n,p}$. A simple way is to "downgrade" the vectors $\vec{E}$ and $\vec{J_{n,p}}$ in 2D or 3D spaces to scalars. For example, the formula implemented in the TCAD application (Cider) used in this study states that[4] the "effective" field for impact ionization is the component

[3] The ideal threshold energy depends on the ratio of the electron and hole conductivity effective mass.

[4] As found in source code file `twoaval.c` [62].

(or vector projection) in the same direction with the current density, i.e.,

$$\text{Field components: } E_{n,p} = \text{proj}_{\vec{J_{n,p}}}\,\vec{E} = \frac{\vec{E}\cdot\vec{J_{n,p}}}{\left|\vec{J_{n,p}}\right|} \tag{2.21}$$

where "proj" stands for vector projection [63], and J_n and J_p are used to calculate the field components in $\alpha_n(E_n)$ and $\alpha_p(E_p)$ for electrons and holes, respectively. This manipulation may be roughly translated into the following. The lucky-drift model requires a minimal threshold energy to be reached for a carrier (electron or hole) to initiate impact ionization. In an infinitesimal space, only the component of $\vec{E}$ that is parallel to $\vec{J_{n,p}}$ can increase the electron's (or hole's) linear velocity and its kinetic energy. In fact, there is a "hidden" term for electric power density $\vec{E}\cdot\vec{J}$ implied in the generation rate $G \propto JE\exp(\dots)$.

Lastly, the unitless ratio $\frac{1}{r}$ in Equation 2.19 needs to be determined for a quantitative evaluation. A commonly used assumption is $\frac{\mathscr{E}_i}{\mathscr{E}_R} = \frac{\lambda_i}{\lambda_R} = r$, so r can be found by measuring $\mathscr{E}_R$ experimentally and using the ideal approximation $\mathscr{E}_i = \frac{3}{2}\mathscr{E}_g$. However, later in Chapter 3, we develop an extraction technique that can determine this value using the terminal characteristics data acquired from our MOSFET I-V test experiment. Therefore, we replace the fraction $\frac{1}{r}$ with an undetermined overall scaling factor A, which may also vary between electrons and holes. Finally,

$$\text{Avalanche rate:} \qquad \alpha_{n,p}(E_{n,p}) = \frac{qAE_{n,p}}{\mathscr{E}_i}\exp\left(-\frac{\mathscr{E}_i}{q\lambda_R E_{n,p}}\right) \tag{2.22}$$

Comments:

1. The "avalanche rate" $\alpha_{n,p}$ in units of $\left[\text{cm}^{-1}\right]$ is mathematically unbounded but physically constrained. In fact, a popular way to evaluate avalanche breakdown, or the scenario when the self-heating at the end of the device's region W exceeds its thermal power capacity, is by evaluating the "multiplication integrals" [22, 64]:

$$1 - \frac{1}{M_n} = \int_0^W \alpha_n \exp\left[-\int_x^W (\alpha_n - \alpha_p)\,dx'\right]dx \tag{2.23a}$$

$$1 - \frac{1}{M_p} = \int_0^W \alpha_p \exp\left[-\int_0^x (\alpha_p - \alpha_n)\,dx'\right]dx \tag{2.23b}$$

where M_n and M_p are the electron and hole current multiplications due to the avalanche

effect (the accumulation of impact ionization over the entire region). Mathematically, M_n and M_p can reach infinity as α increases, and the two integrals on the right hand side can be as large as 1. However, it is practically impossible due to the device's thermal capacity. As external stress increases, the avalanche effect causes more and more thermal energy dissipation, and eventually the device will be destroyed by the self heating effect. In practical applications, the avalanche breakdown evaluation usually sets the integral threshold just below unity, or when the total current exceeds the designed thermal capacity [65]. The breakdown voltage is defined as the lowest value that makes the integrals reach the threshold. Regardless, assuming $\alpha_n = \alpha_p$, one may get

$$\int_0^W \alpha_{n,p} dx \leq 1 \tag{2.24}$$

and the equality is true when the device experiences unbounded impact ionization and breaks down, highlighting the probabilty density's aspect of the avalanche rate $\alpha(E)$.

2. An electron is not obligated to lose all its gained energy in one phonon scattering event. It can continue to travel with its remaining energy, and can also gain more energy, though the possibility is low, considering the scattering angle is a random variable. The multi-stage process [66] includes this factor and gives a similar but more complicated result for the avalanche rate. In fact, the result in the referenced work will become equivalent to Equation 2.22 under the assumptions of low field ($E \lesssim 1 \times 10^5\,\mathrm{V/cm}$) and low probability of impact ionization versus phonon scattering.

3. The lucky-drift model reduces the calculation complexity of the impact ionization rate by explicitly relating it to the phonon scattering rate, which can be determined in advance, such as by combining experimental results [67]. As will be seen in Section 2.3, the local field can be estimated by using applied terminal voltages, enabling circuit-level simulations using only lumped components. In contrast, as stated before, one can directly find λ_i, such as by Monte Carlo simulation [58, 59], which looks at the full picture of electron transport. Under applied fields, the electron gas deviates from the equilibrium state, and tends to have a larger "high-energy tail" (population of electrons

with energy above the threshold $\mathscr{E}_i$) than calculated with the Maxwellian in Equation 2.2. The avalanche rate α can be determined by counting the impact ionization events per unit time. While this method provides more physical insights, it requires additional complexity in the calculation.

4. Since our goal is to efficiently simulate multi-device circuits, we use a compact model that we create and calibrate consisting of only closed-form expressions and depending only on terminal inputs. The avalanche rate derived in this section, $\alpha(E)$ in Equation 2.22, only depends on the local field E, which is approximated in terms of terminal voltages in Section 2.3. All other parameters are determined ahead of circuit simulations, by either experimental or simulation data. As described in Chapter 3, we use the data from I-V measurements that we performed for the compact model extraction.

2.2 Device-Level Modeling of Impact Ionization in N-type MOSFET

We use the device-level simulation to provide insights to the device's internal behavior when impact ionization and parasitic bipolar currents are present. With the simulation data, the physical-level model can be calibrated and then reduced into circuit level models, as shown previously in Figure 2.1. An n-type silicon MOSFET device is simulated using Cider [68], which is an open-source 2-D device equation solver integrated with CoolSpice [69], which is a circuit simulation package provided for free for student use. The transistor structure is based on $0.18\,\mu$m planar technology.

2.2.1 Device-Level Simulation of Planar Silicon MOSFET

In the device-level simulation, the following set of equations, or the Drift-Diffusion equations, are solved self-consistently.

$$\text{Poisson's Equation:} \quad \nabla^2\phi = \frac{q}{\varepsilon}\left(N_D - N_A + n - p\right) \tag{2.25a}$$

$$\text{Electron Current Continuity:} \quad \frac{\partial n}{\partial t} = \frac{1}{-q}\left(-\nabla\cdot\vec{J}_n\right) + G_n - R_n \tag{2.25b}$$

$$\text{Hole Current Continuity:} \quad \frac{\partial p}{\partial t} = \frac{1}{+q}\left(-\nabla\cdot\vec{J}_p\right) + G_p - R_p \tag{2.25c}$$

$$\begin{aligned}\text{Electron Current Components:} \quad \vec{J}_n &= (-q)\ n\ (-\mu_n)\ (-\nabla\phi) \quad \text{(drift)} \\ &\quad +(-q)\ D_n\ (-\nabla n) \quad \text{(diffusion)}\end{aligned} \tag{2.25d}$$

$$\begin{aligned}\text{Hole Current Components:} \quad \vec{J}_p &= (+q)\ p\ (+\mu_p)\ (-\nabla\phi) \quad \text{(drift)} \\ &\quad +(+q)\ D_p\ (-\nabla p) \quad \text{(diffusion)}\end{aligned} \tag{2.25e}$$

where ϕ is the electric potential, n is the mobile (in the conduction band) electron concentration, $\vec{J}_n$ is the electron current density, μ_n is the electron mobility, and D_n is the electron diffusivity. G_n and R_n are the electron generation and recombination rate. p and all subscriptions p stand for the mobile (in the valence band) holes. N_D and N_A are donor and acceptor doping concentration. $q = +1.602\cdots\times 10^{-19}\,\text{C}$ is the electron charge magnitude.

ε is the material (silicon or SiO_2) dielectric constant.

All these variables are local variables dependent on position and time. It means we can create a set of equations describing an actual device with an almost real geometric shape and using a set of distributed parameters. Currently, the device structure used in this work is a planar N-type MOSFET based on the 0.18 µm process. The design parameters are carefully chosen; our device-level model can always be adapted to a MOSFET design of interest. The critical parameters used in the device simulation in this work are listed in this section. Geometric dimensions are illustrated in Figure 2.2 and elaborated in Table 2.1. The gate overlap refers to the small overlapping distance between the gate and the lateral diffusion of the self-aligned source and drain wells introduced by the annealing process.

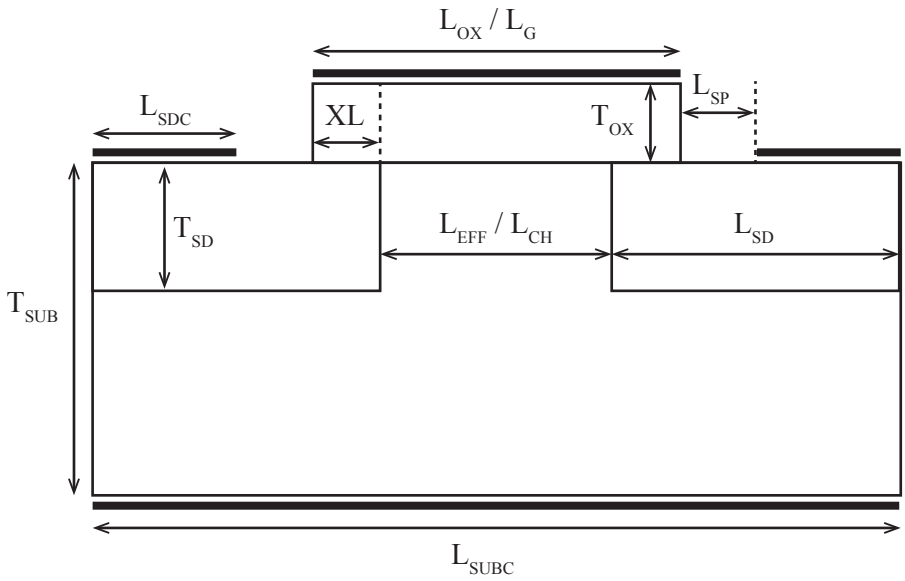


Figure 2.2: Schematic of a simulated MOSFET in this work. Critical dimensions are labeled, and the values are given in Table 2.1. Terminal contacts are presented as thick lines.

Table 2.1: Geometric dimensions of simulated MOSFET, as illustrated in Figure 2.2

Critical dimensions		Value [nm]
T_{OX}	Gate oxide thickness	4.1
L_{CH}	Channel length (effective)	180
L_{SD}	Source / drain lateral length	200
T_{SD}	Source / drain implant depth	100
XL	Gate overlap	10
L_{OX}	Gate metal and oxide length	$2XL + L_{ch}$
Non-critical dimensions		
T_{SUB}	Substrate thickness	> 200
L_{SP}	Sidewall spacer thickness	100
L_{SDC}	Source/drain metal contact length	200
L_{SUBC}	Substrate metal contact length	$2L_{SD} + L_{CH}$

The doping profile parameters entered into the device simulator are listed in Table 2.2. An illustration of the final doping profile is shown in Figure 2.3. The extracted gate threshold voltage $V_{TH} = 0.85\,\mathrm{V}$. The built-in voltage of the source-body junction is found using data in Table 2.2 as $0.99\,\mathrm{V}$.

Table 2.2: List of doping profile parameters of the MOSFET in simulation

Region (Dopant type)	Distribution	Parameter	Value
Substrate (p)	Uniform	Concentration	$1 \times 10^{18}\,\mathrm{cm}^{-3}$
Drain and Source (n)	(Vertical) Gaussian	Peak Concentration C_{peak}	$2 \times 10^{20}\,\mathrm{cm}^{-3}$
		Depth R_p	20 nm
		Char. length $\sqrt{2\left(\Delta R_p^2 + Dt\right)}$	40 nm
	(Lateral) Erfc	Diffusion char. length $\sqrt{Dt}$	6 nm
Gate poly-Si contact (n)	Uniform	Concentration	$1 \times 10^{19}\,\mathrm{cm}^{-3}$

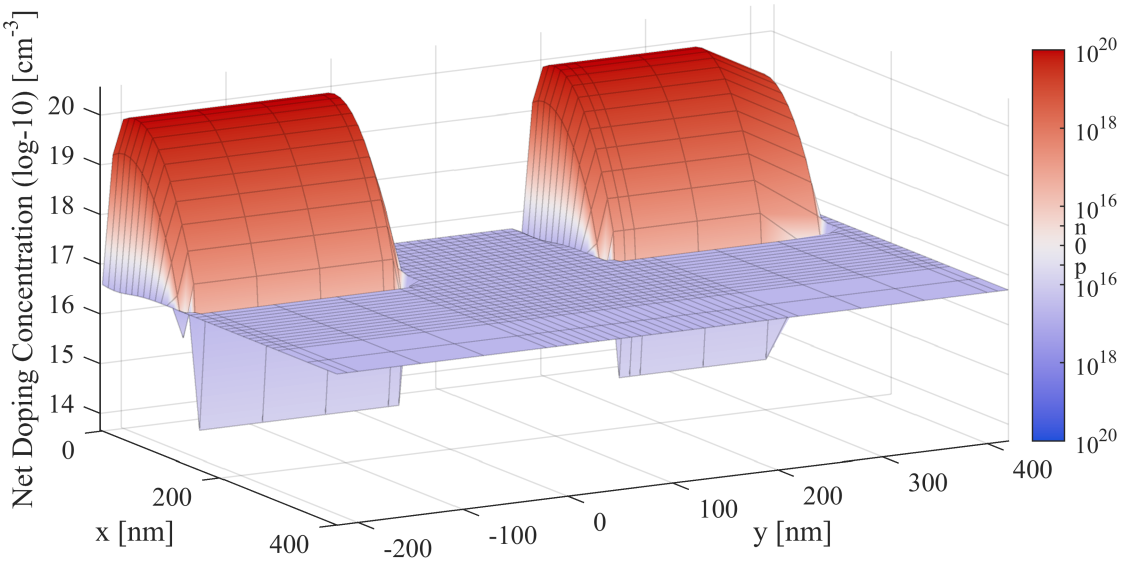
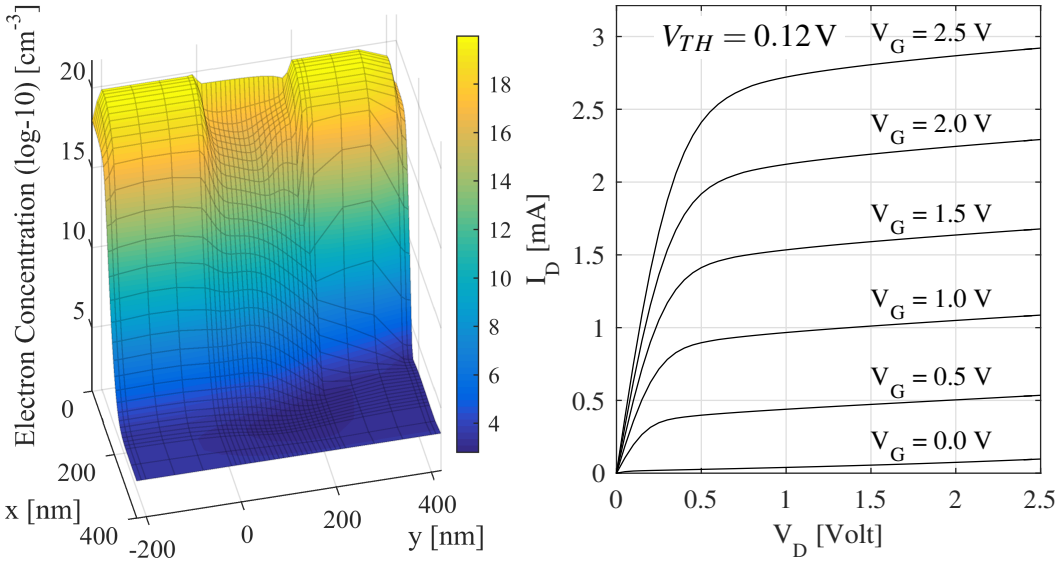


Figure 2.3: Plot of net dopant concentration in the simulated MOSFET device. The channel starts at $(0,0)$. Parameters for doping profile entered into the simulation program are listed in Table 2.2.



(a) Device electron concentration (b) I-V curves

Figure 2.4: A basic simulation of the numerical MOSFET device as described in Section 2.2.1. Threshold voltage $V_{TH} = 0.12\,\mathrm{V}$, channel width $W = 1.4\,\mu\mathrm{m}$. Source and body contacts are grounded.
(a) Electron concentration plot inside the device, $V_G = V_D = 2\,\mathrm{V}$.
(b) I-V curves of a DC voltage sweep.

Shown in Figure 2.4 is a basic simulation (free from secondary carriers) of the MOSFET device with threshold voltage $V_{TH} = 0.12\,\text{V}$ and channel width $W = 1.4\,\mu\text{m}$. The source and body contacts are grounded. The transistor is tested using DC sweep simulation. The gate voltage is fixed at various levels, and the drain voltage is swept while the drain current is recorded. This calculation only captures the "standard" drift and diffusion current without impact ionization (the generation terms in Equation 2.25b and 2.25c), although the field may be high enough to cause impact ionization. Once the transistor has entered the saturation region, the current stays fairly constant. The slight increase following the increase in drain bias is due to channel length modulation and drain-induced barrier lowering. This data will be used as a baseline for modeling impact ionization in the rest of this study.

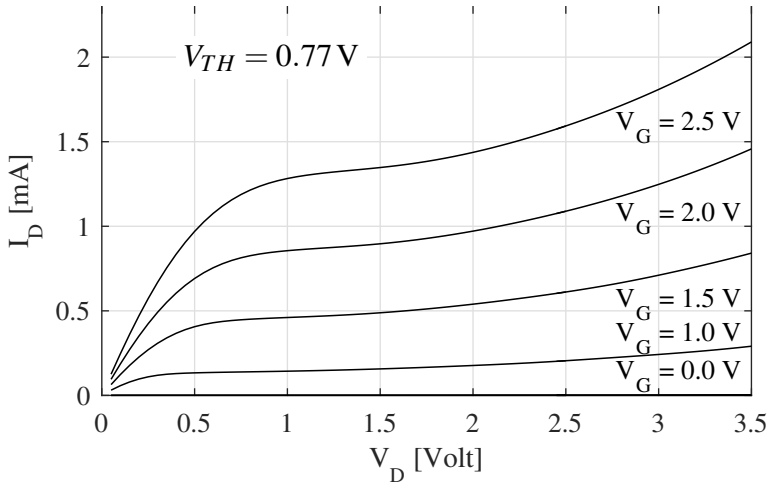


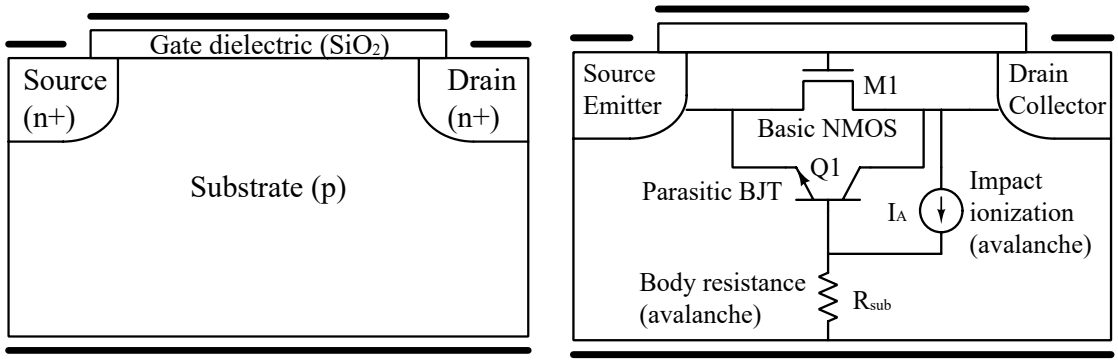
Figure 2.5: I-V curve of device-level simulation with impact ionization calculation. $V_{TH} = 0.77\,\text{V}, W = 1.4\,\mu\text{m}$. Due to a different device setup, the current readings in the low-voltage range are different than in Figure 2.4. More explanation is in the text.

Another simulation is performed with local impact ionization calculation with a different $V_{TH} = 0.77\,\text{V}$ and $W = 1.4\,\mu\text{m}$. The I-V curves are shown in Figure 2.5. Under very low gate bias ($V_G = 0.5\,\text{V}$ and $1.0\,\text{V}$) the drain current is minimal. Under very high drain bias conditions, the drain current is significantly increased, in contrast to the basic simulation as shown in Figure 2.4. A different device was used to produce this example result. Compared to the device used in Figure 2.4, the substrate doping was changed from

$5 \times 10^{16}\,\mathrm{cm}^{-3}$ to $1 \times 10^{18}\,\mathrm{cm}^{-3}$ causing an increase in the threshold voltage. Therefore, the drain current is generally lower for the same applied gate bias, which is more obvious in the low-drain-voltage regions, where the impact ionization current is negligible. However, the influence of impact ionization in the high-drain-voltage region is still obvious.

2.2.2 Parasitic Bipolar Behavior in N-type Silicon MOSFET

In a regular n-type MOSFET, the source-substrate-drain structure is essentially an n-p-n-doped semiconductor, which resembles an npn bipolar junction transistor (BJT). Effectively, there is a parasitic bipolar structure in the N-MOSFET. Under normal operation conditions, this parasitic BJT is not under forward active mode, contributing minimal to none to the MOSFET's behavior. However in this section, we will discuss the abnormal condition when the impact ionization-generated current activates this parasitic structure and the additional undesired consequences.



(a) Basic planar N-MOSFET geometric structure

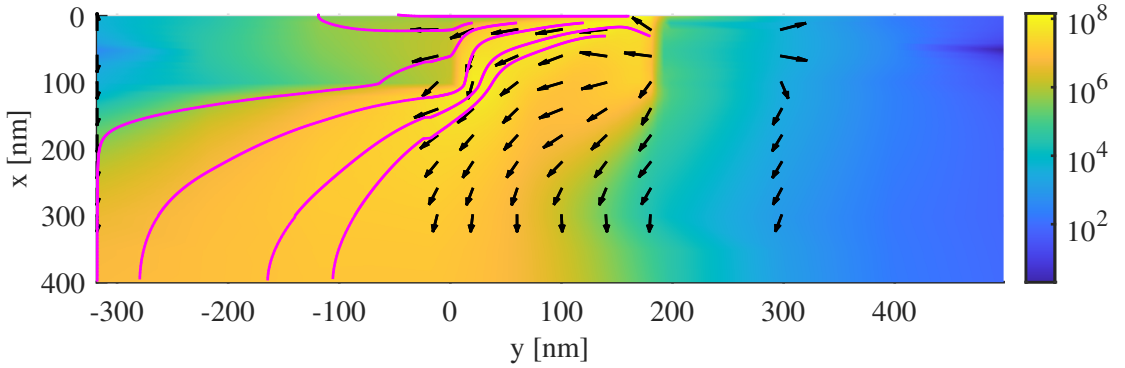
(b) Equivalent circuit showing the N-MOSFET and parasitic structures

Figure 2.6: Illustrations of a planar N-MOSFET and the parasitic bipolar structure

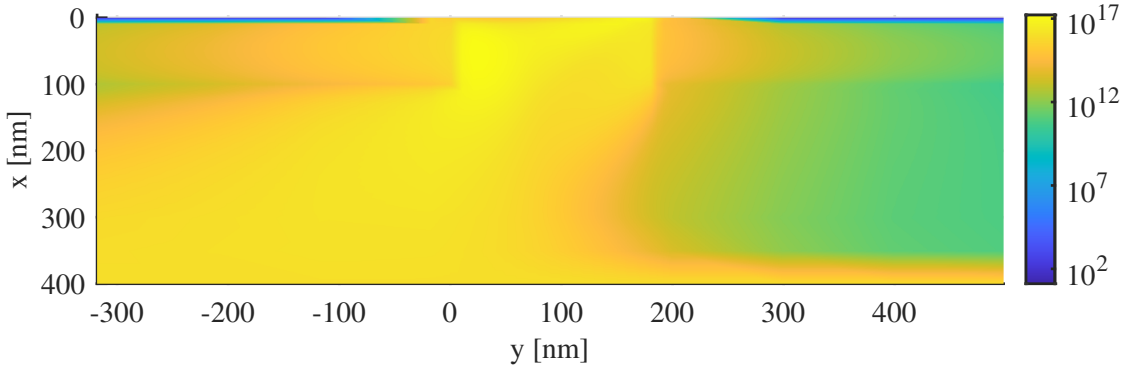
As illustrated in Figure 2.6, we can describe this phenomenon with a bipolar junction transistor (BJT) in parallel with the MOSFET [26], sharing the source as its emitter, the substrate as its base, and the drain as its collector. The basic structure including the gate dielectric (SiO_2), source and drain wells (n+ doped), and substrate (p doped) are shown in Figure 2.6a. The four contact terminals are illustrated with solid lines. Next, an equivalent circuit schematic is drawn in Figure 2.6b on top of the device structure graph, showing the basic MOSFET component (M1), the parasitic BJT originating from the n-p-n strcture (Q1), and the current source (I_A) for avalanche current due to channel impact ionization, along with its associated effective resistance (R_{sub}) creating a forward potential bias for the

BJT from the impact ionization-generated hole current.

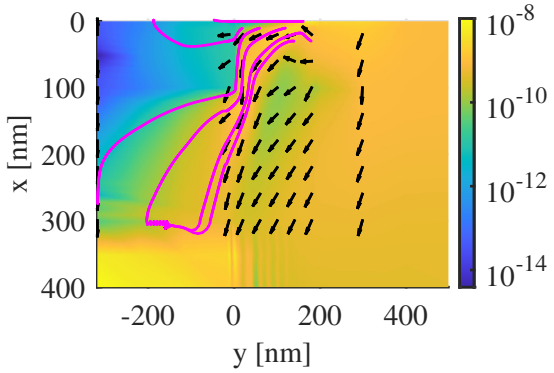
In Figure 2.7 in the next several pages, simulation results are shown for impact ionization calculation included and not included for comparison. The device setup is slightly different than before as in Figure 2.3. Namely, the source and drain (n+) implant regions are intentionally made with abrupt doping profiles so that the source-body and drain-body interfaces can be clearly located at $x = 0\,\text{nm}$ and $x = 180\,\text{nm}$, respectively. The uniform n+ doping depth is $0 \leq y \leq 100\,\text{nm}$. The gate region is $x \leq 0\,\text{nm}$ on top of the channel and the interface and is hardly seen in the figures. Other the critical geometric dimensions remain the same as in Section 2.2.1.



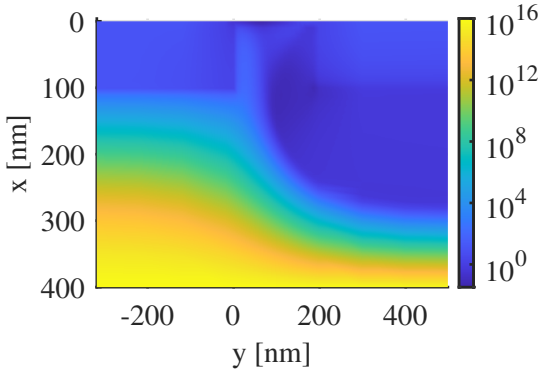
(a) Hole current density J_p with impact ionization [mA/cm^2]



(b) Hole concentration p with impact ionization [cm^{-3}]

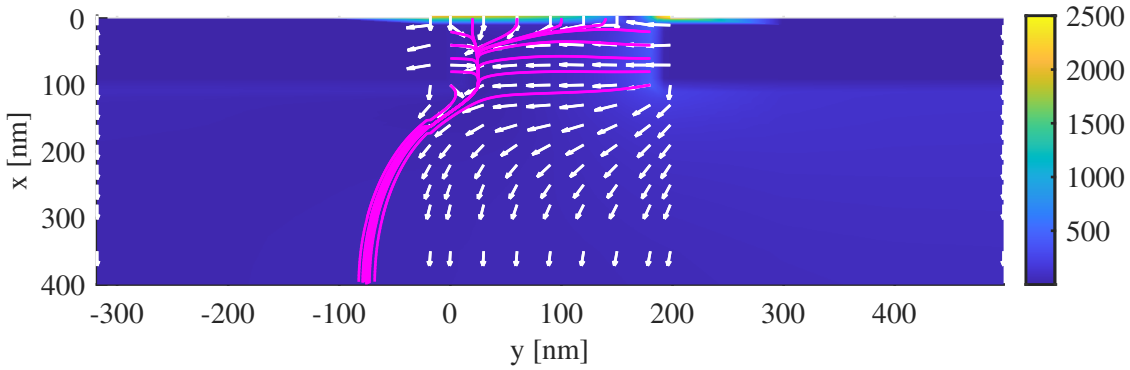


(c) Hole current density J_p without impact ionization [mA/cm^2]

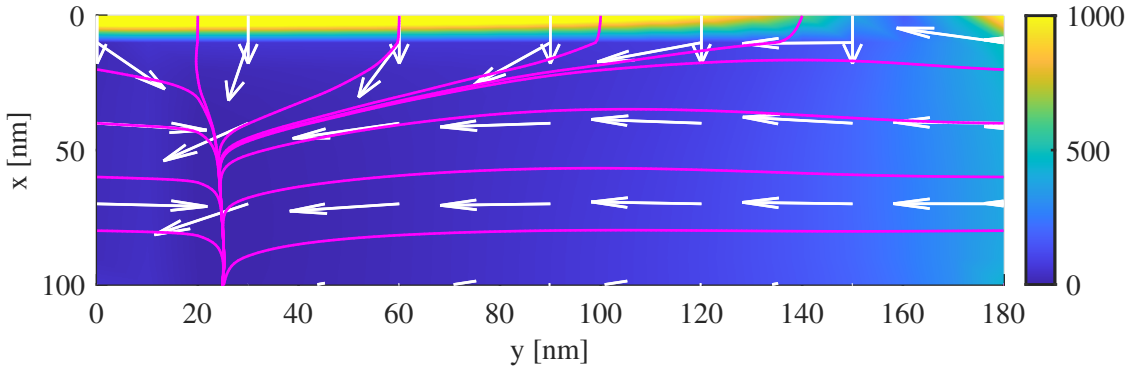


(d) Hole concentration p without impact ionization [cm^{-3}]

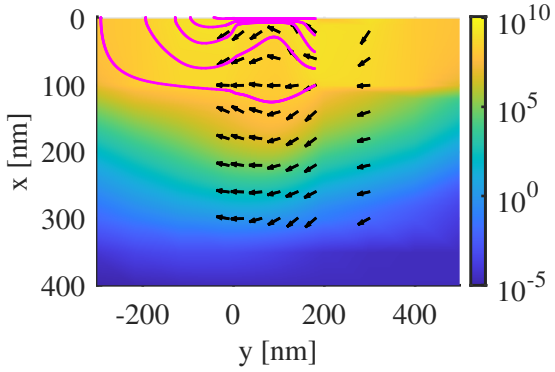
Figure 2.7: Device data with impact ionization generation rate calculation, and without it for comparison. In this page: hole current density with impact ionization (a) and without (c), and hole concentration with (b) and without (d).
The device setup is similar to formerly used in Figure 2.5. More explanation is in the text. Bias $V_{GS} = V_{DS} = 2.00\,\mathrm{V}$. Colors show the magnitude, arrows show vector directions, and streamlines trace the field.



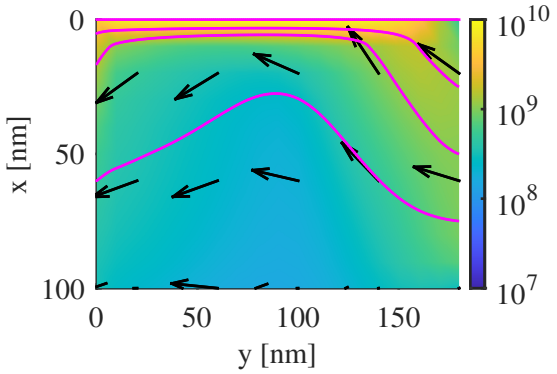
(e) Electric field E with impact ionization [kV/cm]



(f) Zoom-in of (e)

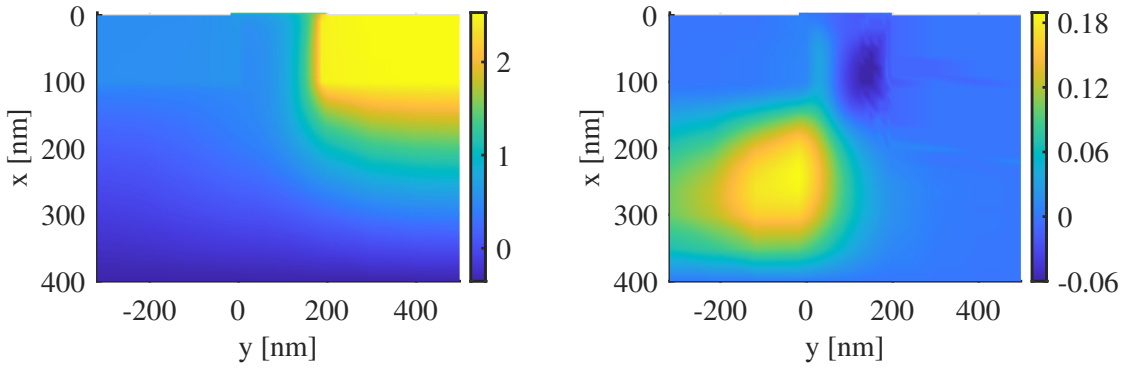


(g) Electron current density J_n with impact ionization [mA/cm^2]



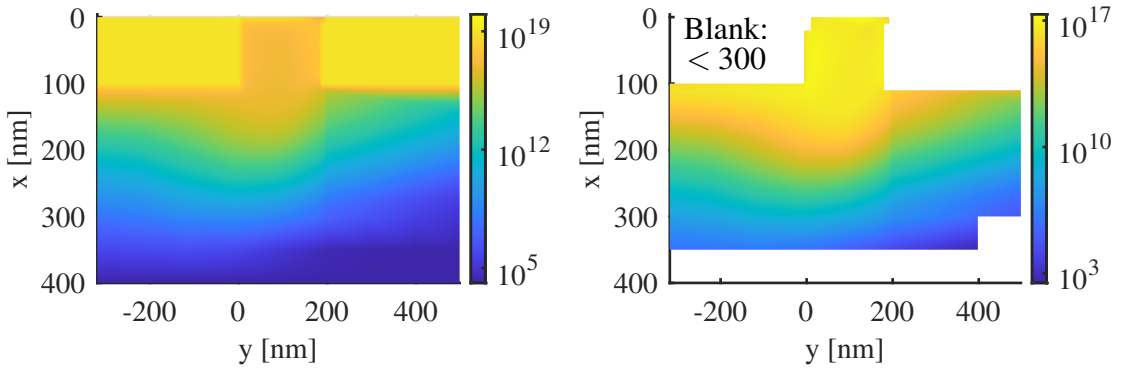
(h) Zoom-in of (g)

Figure 2.7: (Continued) Device data with impact ionization generation rate calculation, and without it for comparison. In this page: Electric field with impact ionization (e), electron current density with impact ionization (g), and zoom-in views (f), (h) for $0 \leq x \leq 100\,\text{nm}$ (below interface) and $0 \leq y \leq 180\,\text{nm}$ (between the source and drain).
The device setup is similar to formerly used in Figure 2.5. More explanation is in the text. Bias $V_{GS} = V_{DS} = 2.00\,\text{V}$. Colors show the magnitude, arrows show vector directions, and streamlines trace the field.



(i) Potential ϕ with impact ionization [V]

(j) Potential change with i.i. $\Delta\phi$ [V]



(k) Electron concentration n with impact ionization [cm^{-3}]

(l) Electron concentration change with impact ionization Δn [cm^{-3}]

Figure 2.7: (Continued) Device data with impact ionization generation rate calculation, and without it for comparison. In this page: potential (relative to intrinsic Fermi level) with impact ionization (i), electron concentration with impact ionization (k), and their relative changes (j), (l) after applying the impact ionization calculation.
The device setup is similar to formerly used in Figure 2.5. More explanation is in the text. Bias $V_{GS} = V_{DS} = 2.00\,\mathrm{V}$. Colors show the magnitude.

Apparently, under normal MOSFET operation, there is no proper biasing to put the BJT under forward (or reverse) active mode. From the MOSFET's perspective, electrons in the source (n+ well) can move freely into the inversion layer at the surface and form the channel. But the potential barrier between the source and the majority of the body, far from the channel, still strongly prohibits any diffusion current between the source and body, which is equivalent to the emitter current in a BJT. Also, there is no significant hole current in the N-MOSFET, since the channel is depleted and inverted, and the reversely biased drain-body junction does not allow conduction.

However, under the exposure to EMI, the holes generated by impact ionization can form a significant current. In Figures 2.7b and 2.7a, the hole concentration and current density in the device simulation including the calculation for impact ionization is shown. In comparison, with no impact ionization calculation, the two hole-related values are minimal as in Figures 2.7c and 2.7d. The generated holes are expelled by the gate-bias field and the built-in field of the source-body junction (shown in Figure 2.7e and 2.7f). Generally in the entire device, $E \leq 400\,\mathrm{kV/cm}$, while it becomes $E \geq 1\,\mathrm{MV/cm}$ at $x = 0$ or the semiconductor-dielectric interface, where the boundary conditions relates it to the oxide field ($\approx 2\,\mathrm{MV/cm}$).

In Figures 2.7g, 2.7h, 2.7k, and 2.7l, the electron concentration and current density substantially increase. Around the interface and channel ($0 \leq x \leq 100\,\mathrm{nm}$), the electron concentration increases by more than 100 %. But more significant increases are found near the source-substrate junction near $x \approx 250\,\mathrm{nm}$ at up to about 1,000 times. Effectively, the "thickness" of the channel is bigger, as more electrons now exist between the drain and source in deeper regions.

Also observed is the increase of the substrate potential near the body-source junction ($\sim 0.1\,\mu\mathrm{m}$ below the metallurgical junction)[5] $\Delta\phi \simeq 0.19\,\mathrm{V}$ as shown in Figure 2.7j. The highest increases in potential can be seen around the source-substrate junction at the sub-

[5]For comparison, in the simulation data based on a realistic device described later in Section 3.2, the reading is $\Delta\phi \simeq 0.26\,\mathrm{V}$ at $\sim 0.5\,\mu\mathrm{m}$ away from the metallurgical junction.

strate (p) side. Typically, one may relate this to the influence of the drain field and describe it with drain-induced barrier lowering (DIBL). However, by examining the simulation data without impact ionization calculation leads us to believe that DIBL alone is not enough to explain the amount of potential increase when impact ionization is significant. Instead, we claim the following.

The generated majority carriers (holes) are immediately "pushed away" by the vertical (gate) and lateral field components and could gather near the body-source junction. But due to the p-n barrier, it is hard for the holes to form a net current into the source, and therefore a local concentration build-up is possible before they reach the body contact. This is similar to the Kirk effect [70, 71] in a BJT, when a high-level injection of majority carriers (electrons) from the base (p) into the collector (n) can cause a local build-up of majority charges in the space-charge region, shifting the junction field and effectively widening the base region. In our case, the amount of generated holes near the source (as high as $\Delta p \simeq 1 \times 10^{17}\,\mathrm{cm}^{-3}$ at $\sim 20\,\mathrm{nm}$ from the metallurgical junction[6]) could partially "shield" the dipoles between the depeleted dopants, shifting the junction field toward the source (n) region and effectively raising the potential on the body (p) side, making it easier for the body-to-source hole current and, more importantly, the source-to-body electron current. In summary, the local body potential depends on the generated hole concentration, which in turn is reflected in the impact ionization current. Additionally, the local potential along the path of the substrate hole current is related to such current by Ohm's law. Therefore, we propose a substrate resistance (R_{SUB}) model that not only has a constant, Ohmic component, but also an additional part depending on the impact ionization current. The augmented substrate resistance model will be introduced in Section 2.3.3.

Finally, the parasitic BJT formed by the drain-body-source regions may be activated by this local potential increase ($\Delta\phi$), effectively acting as the base-emitter forward bias in a typical BJT circuit. The additional emitter (or source) current due to the BJT's activation

[6] In the device in Section 3.2, the reading is $\Delta p \simeq 1 \times 10^{15}\,\mathrm{cm}^{-3}$ at $\sim 0.5\,\mu\mathrm{m}$ from the metallurgical junction

can substantially increase both the source and the drain current.

Furthermore, it is possible that the parasitic BJT current could cause the channel field to collapse, effectively reducing the channel resistance and lowering the drain terminal voltage, similar to a self-activated silicon controlled rectifier (SCR) device. This is called the "Snapback" phenomenon after the high-current, low-voltage region in the I_D-V_D curve, observed in Figure 3.3 in Chapter 3.1. In the compact circuit model derived in Chapter 2.3, the BJT in parallel to the MOSFET provides a possible second solution to the circuit *I-V* relationship, which cannot be achieved by a single compact MOSFET component because commonly used SPICE models (e.g. BSIM and SPICE Level 1-3) are "I-V" models defined by the one-way dependency of terminal currents on terminal voltages. Since the BJT current can sustain itself without the "regular" MOSFET channel current controlled by the gate field, it is possible that, under extreme conditions, the MOSFET device (including the parasitic structure as a whole) becomes trapped in this low-resistance state even after the positive V_{GS} is removed.

2.3 Circuit-Level Modeling of EMI-induced Vulnerabilities in N-type MOSFET (The Soft Error Model)

So far, we have discussed the Snapback phenomenon in a singular MOSFET. It includes the exponential current increase when the transistor experiences a high drain terminal voltage and large internal field as well as the additional current increase due to a parasitic bipolar structure being activated under such conditions. When the stress in the terminal voltage is extremely high, it is also possible that the internal current becomes uncontrollably large, and the internal field collapses due to the unusually high conductance, so the effective resistance suddenly decreases. The device (and the circuit) may fall into a trapped state with a non-monotonous solution to the I-V relationship, generating erroneous responses. All the above behaviors may lead to or intensify Soft Errors in MOSFET circuits, such as signal distortions and bit errors.

Now, our goal is to recreate the Snapback-related behaviors in practical MOSFET circuits in SPICE simulations, and hence we will be able to evaluate a circuit's vulnerability — and reliability — under the influence of transient EMI.

To achieve this goal, we desire a compact model for the device-level vulnerability behaviors for the following reasons. First, device-level simulations solve the distributed, space-dependent system of equations (Equation 2.25) self-consistently and iteratively for internal states (e.g., electron concentration, potential, etc.). This is a time-consuming calculation process that only grows more complex as the circuit's size becomes larger, containing more and more components. Second, the device-level local generation rates due to impact ionization (Equation 2.20) contain highly non-linear functions. Under stress conditions, the exponential terms vary rapidly in the drift-diffusion system of equations (2.25 and 2.20) and may become unstable when improperly conditioned, leading to convergence difficulties. Therefore, it is impractical for a circuit designer to use a device simulator to evaluate a functional circuit's vulnerability when the resource-demanding Snapback cal-

culation is included; because of the extreme non-linearity of local ionization, it becomes prohibitively expensive to use.

On the other hand, a compact model represents a lumped circuit component for its *terminal* voltage-current relationships using only closed-form functions. Time dependence and small-signal frequency response may also be included. Moreover, as will be seen in this section, the compact model can also include physical parameters such as device geometrical dimensions so that a circuit designer may adjust the device design, which is automatically scaled, as long as the model is calibrated.

In this section, we develop the "Soft Error model" — a compact model for the Snapback-related vulnerabilities for a planar N-MOSFET. A schematic for the model's equivalent circuit has been introduced before in Figure 2.1 (and again in Section 2.3.3). We apply our device-level knowledge and our custom I-V relationship equations to form an equivalent circuit [27] representing the MOSFET's vulnerable aspects.

First in Sections 2.3.1 and 2.3.2, we approximate the local impact ionization phenomenon with an overall "avalanche" effect, and develop an avalanche current model (I_A). Disruptions in the power lines can substantially increase the drain-source voltage (V_{DS}) for a short time. Increasing V_{DS} will increase the lateral field in the MOSFET channel and the drain-body space charge (or pinch-off) region. The impact ionization in the channel under substantial lateral (channel) field leads to an exponential growth in drain (ΔI_D) and body current (ΔI_B).

Next in Section 2.3.3, we summarize the parasitic bipolar structure with a BJT component (Q1) and an effective substrate resistance (R_{SUB}) from observations of the device-level simulation data. Holes generated by the impact ionization events and transported to the body contact create an additional field around the source well and an associated potential, which can forward bias the body-source junction. The parasitic source current (ΔI_S) due to the drain(n)-body(p)-source(n) BJT structure is activated by such impact ionization-induced body (hole) current.

For the basic MOSFET component (M1), we choose popular SPICE models such as SPICE Level-3 (from SPICE version 3f5 [72]) and the planar models from the BSIM family [21].

Finally, to conclude this chapter, we use our proposed "Soft Error model" and perform a simple time-dependent simulation for a CMOS inverter in SPICE, showing a potential bit error under EMI disruptions. In Chapter 3, we will perform experimental tests and collect calibration data, and then extract realistic model parameters for an N-MOSFET test device. We will also use this calibrated model to analyze the vulnerabilities of several example circuits under EMI disruptions.

2.3.1 Avalanche Rate at Circuit Level

Now, we adapt the current continuity equations for electrons and holes (same as Equations 2.25b and 2.25c)

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \vec{J}_n + G_n - R_n \tag{2.26a}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \vec{J}_p + G_p - R_p \tag{2.26b}$$

to an N-MOSFET channel. The divergence operator $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)$ can be approximated and simplified to a 1D derivative $\frac{\partial}{\partial x}$ for the MOSFET channel current path (it generally flows from the source to the drain). The $(G-R)$ terms represent local generation and recombination. For typical MOSFET devices with channel lengths around or shorter than 1 µm, it is safe to ignore the recombination effect $R_{n,p} \approx 0$, since the average carrier life time is much longer than the carrier "time of flight" or the time it takes for an electron to drift through the channel (typically $\sim$10 ps for a 1 µm channel).

The generation terms $G_{n,p}$ containing the impact ionization rate are defined in Equation 2.20 as

$$G_n = G_p = \frac{1}{q}\left|\vec{J}_n\right| \alpha_n(E_n) + \frac{1}{q}\left|\vec{J}_p\right| \alpha_p(E_p) \tag{2.27}$$

where $\left|\vec{J_{n,p}}\right|$ are the magnitudes of the local electron and hole current density vectors, and

$E_{n,p}$ are the components of the $\vec{E}$ field parallel to $\vec{J_{n,p}}$ defined in Equation 2.21. In an N-MOSFET channel, we can neglect the hole-induced impact ionization $\alpha_p(E_p)$ since electrons are still the majority carriers. Following this, we may omit the subscripts "n" in $\alpha_n(E_n)$ unambiguously. Also, after the simplification done in Equation 2.21, the E-field $\vec{E}$ and current density $\vec{J_{n,p}}$ are reduced to scalars. Therefore, from now on, we use the simplified notations E for the "scalar" E-field and $J_{n,p}$ for the "scalar" current densities.

In devices with relatively long channels compared to average distance electrons travel without scattering (or the mean free path; usually $\sim$ 10–20 nm in Si), the impact ionization rate $\alpha(E)$ is a function of E. We use the formula derived in Section 2.1 in one dimension, and

$$\alpha(E) = \frac{qAE}{\mathscr{E}_i} \exp\left(-\frac{\mathscr{E}_i}{q\lambda_R E}\right) \tag{2.28}$$

which the same as Equation 2.22.

We assume the steady-state condition here, or $\frac{\partial n}{\partial t} = \frac{\partial p}{\partial t} = 0$, because the time scale of the signals and interferences in the circuit is much longer than the "time of flight" for electron transport. Therefore, Equation 2.26b can be simplified to

$$\begin{aligned} \frac{dJ_n}{dx} = -qG_n &= -\alpha(E) J_n \\ &= \frac{qAEJ_n}{\mathscr{E}_i} \exp\left(-\frac{\mathscr{E}_i}{q\lambda_R E}\right) \end{aligned} \tag{2.29}$$

where the negative sign is associated to the direction of the current, after Equations 2.27 and 2.28 are substituted for G_n and $\alpha(E)$.

Our interest is in the accumulation of local impact ionization events along the MOSFET channel (electron current or J_n for an N-MOSFET), that is the avalanche effect. With the MOSFET source on the left and drain on the right, J_n flows to the $(-x)$ direction, but the avalanche effect grows in the $(+x)$ direction. We use I_{n0} for the "initial" or the "regular" channel current, which can be found near the source (before impact ionization starts to accumulate) and using the low-field MOSFET channel current calculation. We use $x \in [0, L]$ for an arbitrary location between the source ($x = 0$) and the drain ($x = L$) where L is the gate or channel length. Then, by integrating Equation 2.29 over the current path (starting

from the source), the avalanche current $I_A(x)$ at location x is

$$I_A(x) = I_{n0}\left(\exp\left[\int_0^x \alpha(E(u))\,du\right] - 1\right) \tag{2.30}$$

and the spatial dependency of the field is emphasized in $\alpha(E(u))$ (u is the dummy variable for the integral).

Finally, we let $x = L$, and apply a first-order approximation to the exponential term, and the overall avalanche current due to impact ionization is given by

$$\text{Avalanche current:} \qquad I_A = I_A(x = L) \approx I_{n0}\int_0^L \alpha(E(u))\,du \tag{2.31}$$

This is an important result since I_A will be the measurable or observable quantity when a MOSFET device is measured in reality. Under stress voltages between the drain and source terminals, the terminal (drain-source) current increases from the theoretical or "regular" value I_{n0} to include the avalanche current and becomes $I_A + I_{n0}$. As will be shown in the following sections, this approximated version of I_A in Equation 2.31 serves as a bridge connecting the device-level model based on semiconductor equations (Equation 2.25) and compact circuit-level models using only closed-form functions of terminal voltages and currents.

2.3.2 Compact Modeling of Impact Ionization

From Equations 2.22 and 2.31, the simplified circuit-level impact ionization current only depends on the electric field inside the channel region. In order to evaluate Equation 2.31 at the circuit level, the spatial integral must be further eliminated, and the local field may be replaced by an overall *effective* field only depending on terminal voltages.

It is possible to accurately calculate the local lateral field $E(x)$ using device-level simulation tools. But for long-channel devices, we can assume that the field in the channel direction inside the drain-body space charge region is much higher than that outside. With the depletion region approximation, the channel field can be found by solving the 1-D Poisson's equation to be linearly increasing in the space charge region near the drain-body junction, and zero elsewhere. However, if the linear field solution is used in Equation 2.22, it would be difficult to solve Equation 2.31 for the total current with avalanche effect[7].

One possible way to alleviate the computational difficulty, for example, is by evaluating the RMS average of the depletion field.

$$\langle E \rangle_{RMS} = \sqrt{\frac{1}{X_{DEP}} \int_0^{X_{DEP}} \left[E_{MAX} \left(1 - \frac{x}{X_{DEP}} \right) \right]^2 dx} = \frac{1}{\sqrt{3}} E_{MAX} \tag{2.32}$$

where the peak field E_{MAX} appears at the metallurgical junction and depends on the applied drain-source voltage V_{DS}. If the total potential drop across the space charge region is ΔV, then the space charge region width from the depletion approximation is

$$X_{DEP} = \sqrt{\frac{2\varepsilon_{Si}}{qN_A} (V_{bi} + \Delta V)} \tag{2.33}$$

where V_{bi} is the built-in potential of the drain-substrate junction, and N_A is the body (or substrate) doping concentration. Also, from simple algebraic manipulations, we have

$$E_{MAX} = \frac{2\Delta V}{X_{DEP}} \tag{2.34}$$

[7]The challenge is finding a closed-form solution of the integral $\int_{x<0}^{0} u \exp\left(-\frac{1}{u}\right) du$.

$$\langle E \rangle_{\mathrm{RMS}} = \frac{1}{\sqrt{3}} E_{MAX} = \frac{2}{\sqrt{3}} \frac{\Delta V}{X_{DEP}} \tag{2.35}$$

Without losing generality, in the following text, the effective E field is chosen to be the average RMS field $\langle E \rangle_{\mathrm{RMS}}$ or E_{RMS}. Other similar quantities can be derived and used in place of E_{RMS} [73], while keeping the form of the field-dependent avalanche rate (Equation 2.22). Again, our current goal is to evaluate the avalanche current I_A (Equation 2.31) which depends on the channel E field.

Finally, we take the important step and eliminate the space dependency in our simulation of the avalanche current. We replace the MOSFET channel E field *on the inside* as well as the spatial integral over the MOSFET channel current path with quantities that we are able to evaluate just using the terminal voltages which can be observed *from the outside* of the device. Thus, spatial-dependent and iterative evaluations are replaced with compact and closed-form expressions, and the calculation of the avalanche current becomes compatible with SPICE circuit simulations.

Assume the resistance in the drain implant region is negligible, and the MOSFET is in saturation, then the total potential drop across the depletion region is $\Delta V = V_D - V_{DSAT}$ where V_D is the applied drain bias, and V_{DSAT} is the drain saturation voltage. Substituting the field into $\alpha(E)$ in Equation 2.22, the integral in Equation 2.31 can be evaluated and becomes

$$
\begin{aligned}
\gamma_A &\stackrel{\text{def}}{=} \int_0^L \alpha\left(E\left(u\right)\right) du \\
&= \frac{qA}{\mathscr{E}_i} \int_0^L E\left(u\right) \exp\left[-\frac{\mathscr{E}_i}{q\lambda_R E\left(u\right)}\right] du \\
&= \frac{qA}{\mathscr{E}_i} \int_{(L-X_{DEP})}^L E_{\text{RMS}} \exp\left[-\frac{\mathscr{E}_i}{q\lambda_R E_{\text{RMS}}}\right] du \\
&= \frac{qA}{\mathscr{E}_i} E_{\text{RMS}} \exp\left[-\frac{\mathscr{E}_i}{q\lambda_R E_{\text{RMS}}}\right] \int_{(L-X_{DEP})}^L du \\
&= \frac{qA}{\mathscr{E}_i} \frac{2\Delta V}{\sqrt{3} X_{DEP}} \exp\left[-\frac{\mathscr{E}_i}{q\lambda_R}\left(\frac{2\Delta V}{\sqrt{3} X_{DEP}}\right)^{-1}\right] X_{DEP} \\
&= \left(\frac{2}{\sqrt{3}}\right) \frac{qA\Delta V}{\mathscr{E}_i} \exp\left[-\frac{\mathscr{E}_i}{q\lambda_R}\left(\frac{2}{\sqrt{3}}\Delta V\right)^{-1} \sqrt{\frac{2\varepsilon_{Si}}{qN_A}\left(V_{bi}+\Delta V\right)}\right]
\end{aligned}
\tag{2.36}
$$

Now that the only unknown is ΔV, which can be determined with device input V_G and V_D, the avalanche current can be determined solely depending on terminal voltages. Eventually, we will use γ_A and write the avalanche current $I_A = I_{n0}\ \gamma_A$. But before this, we need to ensure that the model equations defined so far are numerically feasible.

It is critical for our simulation model to be within the capability of the simulation tool (SPICE) that evaluates it. We want the model to be "well-behaved" for a wide range of input conditions. One challenge is determining the drain saturation voltage V_{DSAT} with a fault-proof method. For example, the textbook model (the SPICE level-1 model) states $V_{DSAT} = V_{GS} - V_{TH}$ when $V_{GS} > V_{TH}$ and $V_{DSAT} = 0$ otherwise. This function has a discontinuity in its derivative and is zero-valued in certain circumstances. These abnormalities may cause computational errors and convergence difficulties. Meanwhile, the depletion region width X_{DEP} should have a limited range between the minimum under forward bias and the full channel length.

Our approach to find ΔV is based on the BSIM3 model smoothing equation [21] and is given as below.

$$\Delta V = V_{DS} - V_{DSATeff} \tag{2.37a}$$

$$V_{DSATeff} = \frac{1}{2}\left[V'_{DSAT} + V_{DS} + \delta v_1 - \sqrt{\left(V'_{DSAT} - V_{DS} - \delta v_1\right)^2 + 4\delta v_1 V'_{DSAT}}\right] \tag{2.37b}$$

$$V'_{DSAT} = V_{DSAT} + \delta v_2 \tag{2.37c}$$

where δv_1 and δv_2 are smoothing factors in Volts, and the quantities in Equation 2.37 above are in Volts. The saturation voltage V_{DSAT} is defined using first principles [74] as

$$V_{DSAT} = \left[V_{GS} - V_{FB} - (\phi_B + \phi_{CH})\right] + \frac{q\varepsilon_{Si}N_A}{\left(C'_{OX}\right)^2} + \sqrt{\frac{q\varepsilon_{Si}N_A}{\left(C'_{OX}\right)^2}\left[2\left(V_{GS} - V_{FB}\right) + \frac{q\varepsilon_{Si}N_A}{\left(C'_{OX}\right)^2}\right]} \tag{2.38a}$$

$$C'_{OX} = \frac{\varepsilon_{OX}}{t_{OX}} \tag{2.38b}$$

$$\phi_B = V_T \ln\frac{N_A}{n_i} \tag{2.38c}$$

$$\phi_{CH} = V_T \ln\frac{N_{CH}}{n_i} \tag{2.38d}$$

where V_{FB} is the flat-band voltage depending on the body (or substrate) and gate metal (or poly-Si) materials, t_{OX} is the oxide thickness, and N_{CH} is the channel electron concentration under strong inversion. Finally, in sub-micron devices where the classical space charge region width (Equation 2.33) could grow to longer than the channel length, it is necessary to regulate X_{DEP} with similar smoothing functions as Equation 2.37.

$$X_{DEP} = \frac{1}{2}\left[X''_{DEP} + x_1 + \delta x_1 + \sqrt{\left(X''_{DEP} - x_1 + \delta x_1\right)^2 + 4x_1\ \delta x_1}\right] \tag{2.39a}$$

$$X''_{DEP} = \frac{1}{2}\left[X'_{DEP} + x_2 + \delta x_2 - \sqrt{\left(X'_{DEP} - x_2 + \delta x_2\right)^2 + 4x_2\ \delta x_2}\right] \tag{2.39b}$$

$$X'_{DEP} = \sqrt{\frac{2\varepsilon_{Si}}{qN_A}\left(V_{bi} + \Delta V\right)} \tag{2.39c}$$

where x_1 is the lower bound of X_{DEP}, and typically $x_1 \approx \frac{1}{3}\sqrt{\frac{2\varepsilon_{Si}}{qN_A}V_{bi}}$. x_2 is the upper bound and is assumed to be approximately equal to the gate length. δx_1 and δx_2 are smoothing factors. All quantities in Equations 2.39a and 2.39b are in centimeters.

During the derivation of the closed-form equations, many approximations are applied. The approximated form of γ_A found in Equation 2.36 can be improved to better match calibration data by adding a few heuristic corrections. We replace the factor $\frac{2}{\sqrt{3}} \approx 1.2$ with an undetermined quantity F_{S1}, and change ΔV to $(\Delta V)^{F_{S2}}$ (where $F_{S2} > 0$) to compensate the discrepancy in the shape of the actual E field and its approximated expression used in Equation 2.32. F_{S1} is unitless, and $(\Delta V)^{F_{S2}}$ is coerced into Volts. These two parameters will be extracted from calibration data (terminal current-voltage characteristics).

Also, we add the parasitic bipolar current $I_{BJT,n}$ to the "regular" channel current I_{n0}, since both components of the current can initiate impact ionization and induce avalanche current. $I_{BJT,n}$ will be modeled next in Section 2.3.3. Finally, combining Equations 2.31, 2.36, the avalanche current I_A becomes

$$\begin{aligned} I_A &= I_{D0}\ \gamma_A \\ &= I_{D0}\ \frac{q}{\mathscr{E}_i}\ (AF_{S1})\ (\Delta V)^{F_{S2}}\ \exp\left[-\frac{\mathscr{E}_i}{q\lambda_R}\ \left(F_{S1}\ \frac{(\Delta V)^{F_{S2}}}{X_{DEP}}\right)^{-1}\right] \end{aligned} \tag{2.40a}$$

$$I_{D0} = I_{n0} + I_{BJT,n} \tag{2.40b}$$

where ΔV and X_{DEP} are defined in Equations 2.37 and 2.39a, respectively.

The initial drain-source current I_{D0} now includes the standard MOSFET current I_{n0} as mentioned in Equation 2.31, as well as the parasitic BJT current $I_{BJT,n}$ to be described shortly. Note that this expression does not include the local E field, but only terminal voltages as variables. Also note that so far, δv_1, δv_2, δx_1, δx_2, N_{CH}, F_{S1} and F_{S2} have yet been quantitatively determined, although their physical meanings have been explained, and their reasonable ranges can be estimated. Later in Chapter 3, they are considered as adjustable model parameters that are extracted from device-specific data using the Genetic Algorithm.

To summarize, the impact ionization behavior of the MOSFET channel is modeled as a dependent current source, as shown in Figure 2.8 besides the regular MOSFET used to

calculate low-field current. More components will be added after introducing the parasitic current components in the next section.

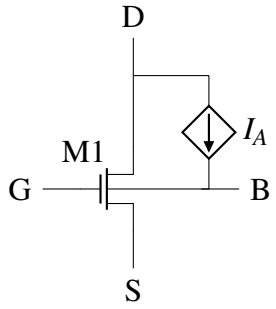


Figure 2.8: Equivalent circuit of MOSFET with avalanche current (I_A) included

Evaluating the ionization rate in terms of terminal voltages as in Equation 2.36 is a general routine in the modeling practice [21, 25, 27, 29], although the dependency may have different and simpler forms. In practical situations, one may simply calibrate the model, including terminal voltages such as V_G and V_D, with experimental data and curve fitting techniques. Analytically solving the Poisson's equation (Equation 2.25a) may yield a position-dependent field, e.g. in a power MOSFET's drift region [75]. Should one need to improve the Soft Error model to better match device-specific characteristics, Equation 2.40 can be modified, while the workflow to evaluate the circuit vulnerability is still easily utilized.

2.3.3 Compact Modeling of Parasitic Bipolar Behavior

As described in Section 2.2.2, due to impact ionization and the parasitic bipolar structure's behavior, there are in total six contributing current components [26], listed in Table 2.3 and illustrated in Figure 2.9.

Table 2.3: Current types in MOSFET used in the circuit model, considering impact ionization and parasitic bipolar structure

Normal MOSFET operation	
I_{ch}	Channel electron drain/source (I_{n0})
Impact ionization generation	
I_{AeD}	Generated electron flowing to drain[1]
I_{AhB}	Generated hole flowing to substrate
I_{AhS}	Generated hole flowing to source
Parasitic bipolar injection	
I_{Pe}	Injected electron flowing from source to drain ($I_{BJT,n}$)
I_{Ph}	Injected hole flowing from channel to source[2]
Terminal total current (externally measurable)[3]	
I_S	Total source current (outbound)
I_D	Total drain current (inbound)
I_B	Total substrate current (outbound)

[1] By applying KCL to the local site $I_{AeD} = I_{AhB} + I_{AhS}$.
[2] Apparently $I_{Ph} = I_{AhS}$ since they are the same current.
[3] Terminal currents I_S, I_D, and I_B are used to build the KCL equation 2.41d.

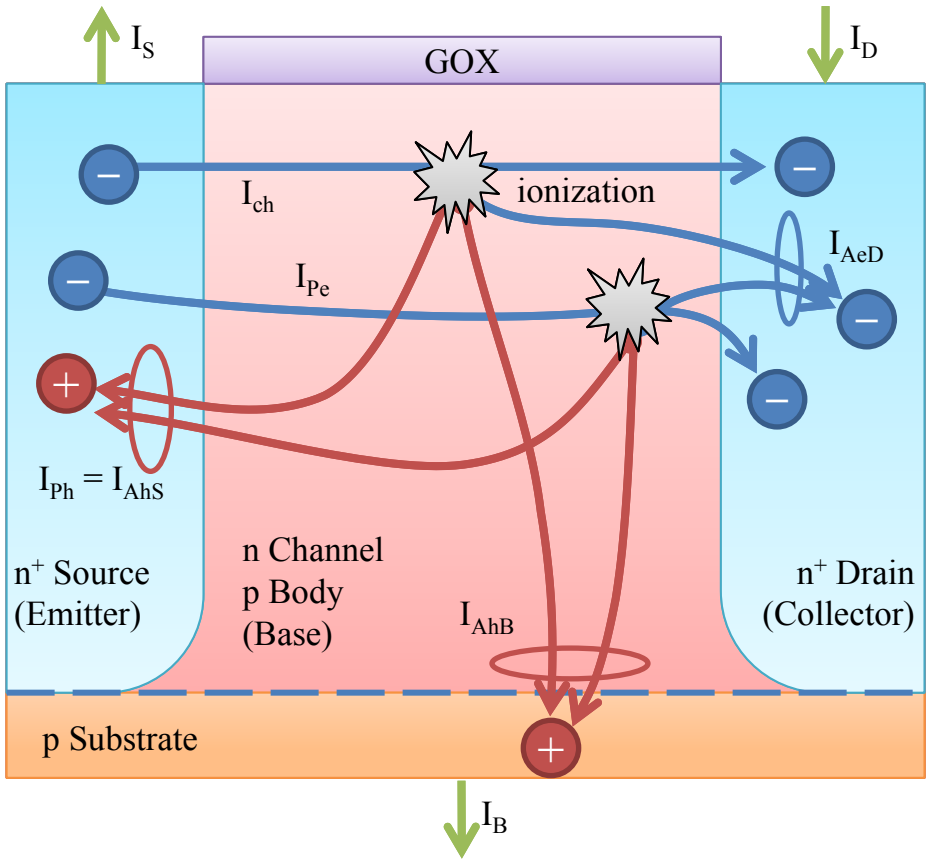


Figure 2.9: Illustration of the current types in an N-MOSFET used in our circuit model, considering impact ionization and the parasitic bipolar structure. The electron current created in the inverted channel under drain bias (typical MOSFET action) as well as in the forward-biased body-source junction (parasitic BJT action) are capable of inducing impact ionization, generating additional electron and hole current. The shapes are not to scale and only indicate existence. The formal description is in Table 2.3.

The normal action of the MOSFET solely involves channel current I_{ch} or "I_{n0}" in Equation 2.40. The remaining five types of currents exist due to impact ionization. Notably, the generated electrons can cause more impact ionization, since they always travel through the drain space charge region before entering the drain.

To further clarify the added internal current components, we write the KCL equations for the inside of device in terms of internal current components

$$\text{Drain Terminal:} \qquad I_D = I_{ch} + I_{AeD} + I_{Pe} \tag{2.41a}$$

$$\text{Source Terminal:} \qquad I_S = I_{ch} + I_{AhS} + I_{Pe} \tag{2.41b}$$

$$\text{Body Terminal:} \qquad I_B = I_{AhB} \tag{2.41c}$$

$$\text{Overall Device:} \qquad I_D = I_S + I_B \tag{2.41d}$$

The three ionization current types with opposite carrier types I_{AeD}, I_{AhS} and I_{AhB} can be further summarized into one "avalanche current" flowing from the drain to the substrate, but the co-existence of electrons and holes does not resemble a p-n junction's dual-carrier behavior under forward bias, because they originate at the same generation sites, and the total avalanche current is equal to *either* the electron *or* the hole current, rather than the sum of all 3 components. At the circuit level, this is equivalent to an internal dependent current source between the two terminals. The "impact ionization current" or "avalanche current" mentioned in Section 2.2.2 is $I_A = I_{AeD} = I_{AhB} + I_{AhS}$.

On the other hand, one can use a lumped BJT circuit component to evaluate the parasitic bipolar current components I_{Pe} and I_{Ph}. The parasitic drain(n)-body(p)-source(n) structure can be treated and modeled as a BJT, but with a different set of model parameters than typical designs. The SPICE BJT Level-1 model is chosen, as it allows us to modify the junction saturation current directly without too much complication. The model equations used in our Soft Error model along with fitting parameters are listed below.

$$\begin{aligned} \text{(Base)} \quad I_{Ph} &= \frac{I_{be1}}{\beta_F} \\ \text{(Emitter)} \quad I_{Pe} &= I_{be1} \\ I_{be1} &= I_{S0}\left[\exp\left(\frac{V_{be}}{N_F\, V_T}\right) - 1\right] \end{aligned} \tag{2.42}$$

where I_{S0}, β_F, and N_F are model parameters defined in SPICE. The above equations are described in manuals of popular SPICE implementations[8]. I_{S0} is the body-source junction saturation current in Amperes (not to be confused with the MOSFET source current I_S), scaled by the effective junction area; V_{be} is the body-source forward potential as will be seen in the circuit schematic Figure 2.10 as V_{Bint}, and in the device-level simulation in Figure 2.7j; β_F is the large-signal (DC) forward current gain; N_F is the junction non-ideality factor.

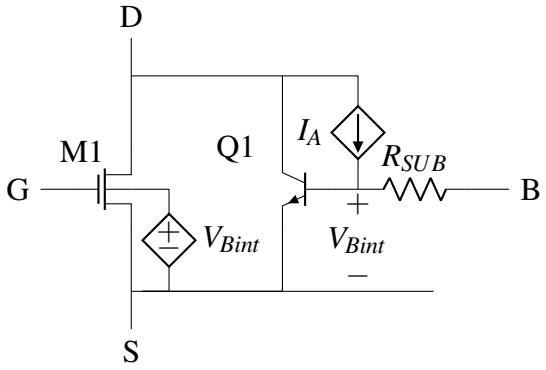


Figure 2.10: Full equivalent circuit for impact ionization and parasitic bipolar behaviors. $M1$ is the original MOSFET without impact ionization. I_A is the avalanche current source. R_{SUB} and $Q1$ are the extracted parasitic bipolar structure. The dependent source V_{Bint} isolates the MOSFET model and the Snapback model to avoid incorrect results (e.g. ionization current flowing back into the MOSFET body). The "internal" body has the increased potential due to the generated hole current, and the body contact sees the "external" terminal voltage.

The full Soft Error model for a single n-type MOSFET is shown in Figure 2.10. M1 is the regular MOSFET model, not containing any impact ionization or parasitic BJT currents. The dependent current source I_A is the avalanche current described in Equation 2.40. R_{SUB} is the effective body/substrate resistance seen by the parasitic BJT. Including

[8]The equations are also listed in the ngspice source code file, `bjtload.c` [62].

the dipole interaction between the generated holes and the space charges from the ionized acceptors described in Chapter 2.2.2, we define

$$R_{SUB} = R_{SUB0} + \frac{a_{di}}{I_A} \ln \frac{I_A}{I_{di}} \tag{2.43}$$

where R_{SUB0} is the linear component from the Ohmic conduction in the substrate The second term ($a_{di} \ln \frac{I_A}{I_{di}}$ in Volts) is the additional potential component originating from the generated carriers which are quantified by the generation (avalanche) current I_A available in circuit-level simulations. a_{di} and I_{di} are proportionality factors to be fitted to measurement data. The second term is limited to be positive using a similar equation to 2.39, introducing two more parameters $v_{di} \gtrsim 0$ (lower bound) and $\delta v_{di} \gtrsim 0$ (smoothing factor). Q1 is the parasitic drain(n)-body(p)-source(n) BJT structure described by Equation 2.42. The optional dependent voltage source V_{Bint} implements the body effect due to the non-zero body-source voltage V_{BS} in the MOSFET model; it is constructed separately so that our additional source I_A does not intefere with the MOSFET model's terminal current evaluation.

The entire circuit is built only using compact device models and closed-form equations. The input variables are limited to terminal voltages and pre-determined parameters. It is possible to treat this circuit as one circuit component representing the MOSFET and its Soft Error behavior. The newly defined terminals D, G, S and B are annotated in Figure 2.10. As later will be seen in Chapter 3.1, the Snapback phenomenon occurs under high V_{DS} when the channel field collapses as the parasitic BJT current dominates, and I_{DS} increases while V_{DS} decreases. This means that the I_D-V_D solution of the MOSFET does not have to be single-valued, and the device could become "trapped" in the region with non-typical solutions.

With initial BJT parameters, a DC simulation is run with a test circuit. The drain current and the gate voltage are driven and controlled, while the drain voltage is recorded and shown in Figure 2.11.

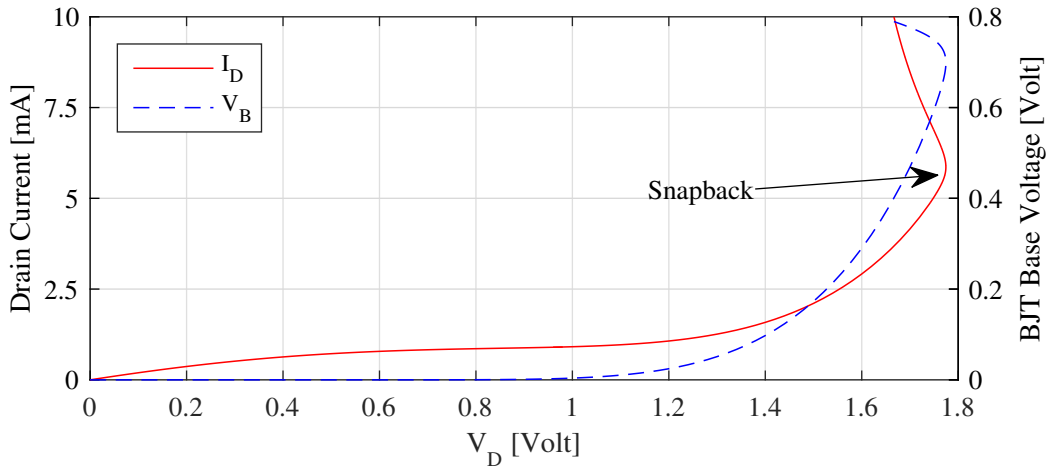


Figure 2.11: I-V curve of simulation of a single n-type MOSFET with impact ionization and Snapback calculation. The total drain current (I_D) and base voltage of the parasitic BJT (V_B or V_{Bint} in Figure 2.10) are shown.
In this example, the Snapback circuit parameters are not yet calibrated. $R_{SUB} = 150\,\Omega$, transistor $\beta = 1$. MOSFET $W \times L = 1.4 \times 0.18\,\mu\text{m}^2$, $V_{TH} \approx 0.8\,\text{V}$. Gate bias $V_G = 2.0\,\text{V}$.

As the driven drain current keeps increasing, the current increase due to impact ionization generation and bipolar diffusion also dramatically increase. At the point when the forward bias condition of the source-substrate junction is enough such that the emitter current becomes comparable to channel current, less drain voltage is required to form the total drain current.

2.3.4 Demonstration of Snapback in an Example Circuit-Level Simulation

So far, we have developed our "Soft Error" model, which is the compact model for impact ionization and parasitic bipolar behavior for a single MOSFET. Now that we have the model equations, it is possible to use estimated model parameters to demonstrate an example circuit simulation. A CMOS inverter circuit, illustrated in Figure 2.12, is simulated in CoolSpice. The test bench includes a matched pair of a P-MOSFET and an N-MOSFET, where the N-MOSFET contains our Soft Error model. We neglect the avalanche and parasitic currents in the P-MOSFET since they are generally less concerning, but one may always apply the same methodology for the P-MOSFET. In this section, the model parameters are not yet calibrated, and estimated values are used. In the next chapter, we will perform experiments and use the measurement data to extract a realistic parameter set.

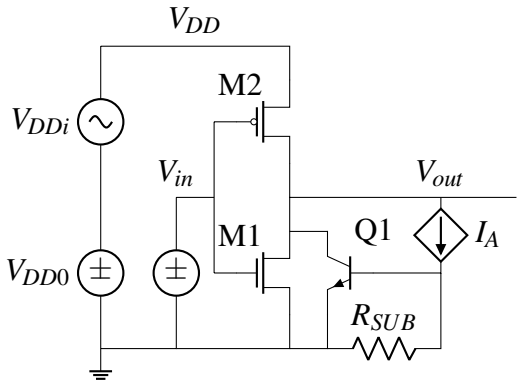


Figure 2.12: A CMOS inverter under test. The impact ionization and parasitic bipolar models are added into the N-MOSFET only, while the P-MOSFET is considered causing less issue because of lower carrier mobility. A pulse voltage source is added into power rail to simulate a transient interference.

In the test bench shown in Figure 2.12, the DC voltage source V_{DD0} provides the power rail $V_{DD} = 2\,\mathrm{V}$ for normal operation. The time-dependent voltage source V_{DDi} simulates disruptions in the power rail caused by external EM interference, following a standard testing methodology for industrial product development and testing [28].

In the bigger picture, a vulnerability specific to CMOS logic circuits called *latch-up* [76] is possible to occur during terminal voltage disruption events. This takes place when

a parasitic p-n-p-n structure inside the CMOS is activated and acts as a silicon controlled rectifier (SCR), creating an undesired parasitic current through the two MOSFET bodies. As a result, the inverter no longer responds to the input (gate) voltage and becomes trapped in a low-resistance state ("latched up").

Unlike the latch-up phenomenon, Snapback can happen in a single MOSFET device. Under a high drain-body field, impact ionization leads to avalanche current and activates the parasitic bipolar structure in a single MOSFET. While it is possible to implement both latch-up and Snapback in a simulation, at this moment, we particularly demonstrate the effects of Snapback only. Generally speaking, latch-up requires a more sophisticated dynamic condition to forward-bias one of the p-n junctions in the SCR (e.g., the output terminal voltage drops below ground or zero), and it can usually be prevented by implanting guard rings to divert the unusual parasitic current [76]. Meanwhile, Snapback only requires impact ionization in the N-MOSFET to occur, which will happen more often.

Time-dependent (transient) tests are simulated under voltage stress on the power rail. An example result is given in Figure 2.13.

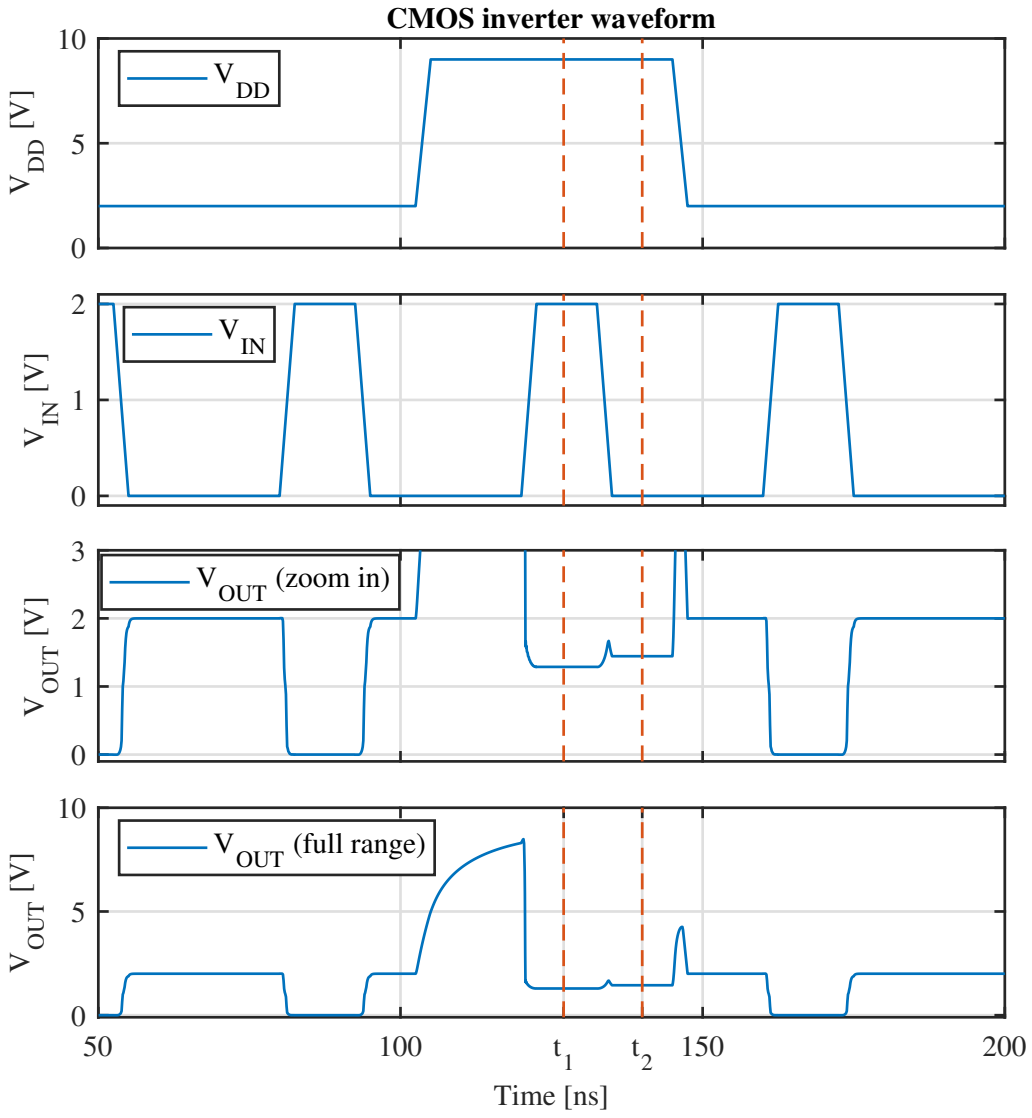


Figure 2.13: Test waveform of the CMOS inverter circuit shown in Figure 2.12. The power rail V_{DD} starts with 2 V, and the input to the gates V_G switches between 0 V and 2 V. A transient interference is added into power rail to temporarily raise it to 9 V. In the waveforms, $t_1 = 127$ ns, $t_2 = 140$ ns.

At time $t_1 = 127\,\text{ns}$, because $V_{DD} > V_{IN} > 0$, the N- and P-MOSFET are both conducting. Due to the impact ionization and parasitic bipolar structure, Snapback occurs, and the N-MOSFET is "trapped" in the Snapback state (high current output with low drain-source voltage). At time $t_2 = 140\,\text{ns}$ when the input voltage falls to low level, the P-MOSFET remains conductive. Meanwhile, in the N-MOSFET, because the parasitic current already exists, and the drain voltage is even higher than at t_1, the total current (regular low-field channel current, avalanche current from impact ionization, and parasitic bipolar current) becomes higher than at t_1.

As a result, both the outputs at the two times are different than expected. For $V_{OUT}(t_1)$, the expected output is 0 V (logic "0"), but the actual output is 1.29 V. The reason for the discrepancy is a mixture of V_{IN} being between the disrupted rails and the Snapback action. For $V_{OUT}(t_2)$, the expected output is at least 2 V (logic "1"), but the actual output is 1.44 V. This result is purely a continuation of the Snapback event. In the context of digital logic, both outputs are within the "undefined" region between ground and power rail. Depending on the design of the digital circuit, this highly suggests a bit error.

Chapter 3: EMI-Induced Soft Error Vulnerabilities: Experimental Tests, Data Extraction, and TCAD Simulations

Our Soft Error model based on Snapback needs calibrating using experimental data. By doing so, we also verify the Snapback phenomenon and the underlying physical mechanisms by observing the MOSFET behaviors under stress. We have conducted experimental DC measurements on individual off-the-shelf N-MOSFET devices (CD4007) that are available from a major electronic device vendor [77] at room temperature. The test results are analyzed in Section 3.1. A device-level model is created using a basic planar Si MOSFET structure to reproduce our experimentally measured terminal characteristics. Through device-level simulation in Section 3.2, the physical aspects of the basic MOSFET behaviors and the Snapback phenomenon are validated. Next, based on our in-house measurement data, the Soft Error model proposed in Chapter 2 is calibrated. The model parameters are extracted using a search method based on the Genetic Algorithm, which will be described in Section 3.3. Finally, practical circuits using the extracted MOSFET and Soft Error models are simulated in the time domain including during transient voltage disruptions in Chapter 4. The simulations show the potential application of the proposed model, and the results further reveal possible EMI-induced Soft Error vulnerabilities when the Snapback phenomenon is present.

3.1 Experimental (DC) Measurements

We developed a measurement technique in-house to measure MOSFET terminal voltages and currents. An HP 4155A Semiconductor Parameter Analyzer ("*the* Analyzer") with four source/monitor units (SMU), two voltage source units (VSU), and two voltage monitor units (VMU) is deployed. We developed an automatic measurement program

mainly in MATLAB which also uses the GPIB driver in Windows to communicate with the Analyzer. The program consists of a couple of hundreds of lines of code and contains two parts. A shared library serves as the intermediate layer between the user and the lower-level GPIB send-wait-receive sequences. It can also easily manage multiple GPIB target devices at once (e.g., the Analyzer along with an additional digital multimeter for real-time monitoring). The user focuses on the actual testing "recipes" and post-measurement data cleaning and consolidation for most of the time, making up the second half of the program.

Our test consists of three phases with different circuit and measurement configurations. A flowchart is shown in Figure 3.1 for the three phases, with the key points for each measurement "recipe" included.

The first two phases are typical N-MOSFET I-V tests. Phase 1 is the I_D-V_{GS} sweep with a low drain voltage $V_{DS} = 50\,\text{mV}$. The gate voltage V_{GS} is driven by one SMU while the drain current I_D is measured by another SMU, which also drives the drain voltage $V_{DS} = 50\,\text{mV}$. The resulting data is a set of (V_{GS}, I_D) sample points. Additionally, the large-signal transconductance G_m is calculated post-measurement given by:

$$G_m = \left.\frac{dI_D}{dV_{GS}}\right|_{V_{DS}=50\,\text{mV}} \tag{3.1}$$

which is then added into the dataset.

Phase 2 is the I_D-V_{DS} sweep with several positive V_{GS} values. The N-MOSFET operates in the "linear/triode" and "saturation" regions with an inversion channel conducting current. The resulting data is a set of (V_{DS}, V_{GS}, I_D) sample points, where V_{DS} is the primary driven variable, V_{GS} is the secondary driven variable, and I_D is the measured variable.

The above two datasets together represent the "regular" N-MOSFET I-V characteristics that resemble textbook equations. A "basic" MOSFET model is extracted from these two datasets, which will be described in Section 3.2.

Phase 3 is the high-voltage and high-current stress test, which measures the Snapback characteristics. Like in Figure 2.11, we expect the N-MOSFET I_D-V_{DS} curves to "bend back" in the high-current regime, leading to secondary drain current (I_D) solutions

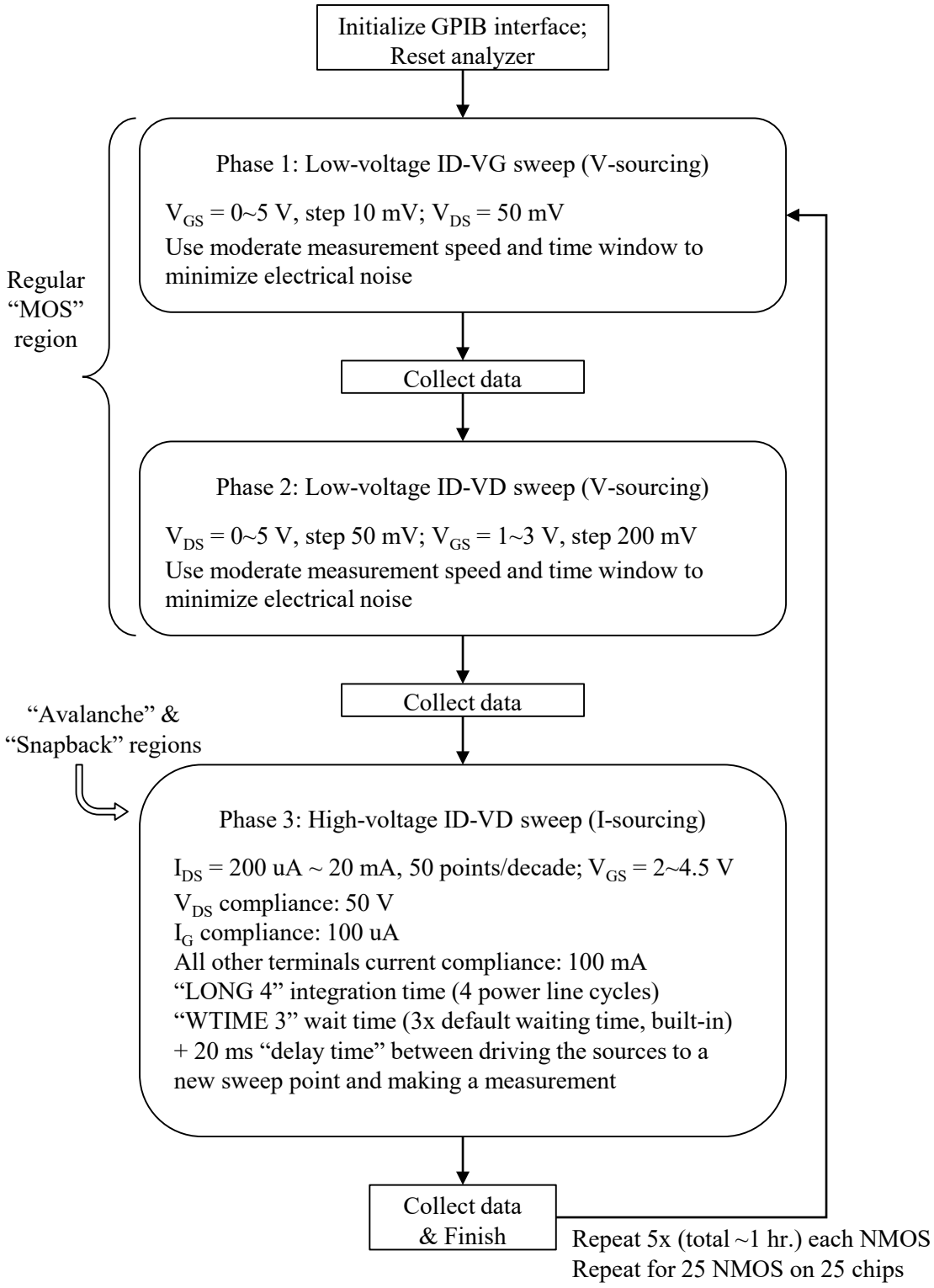


Figure 3.1: Workflow for our Soft Error characteristic I-V measurement. The "MOS", "Avalanche", and "Snapback" regions refer to the I-V data in Figure 3.3.

to the same terminal voltages (V_{DS},V_{GS}). Therefore, we set the SMU connecting to the drain terminal to current-source mode, and I_D is the driven variable.

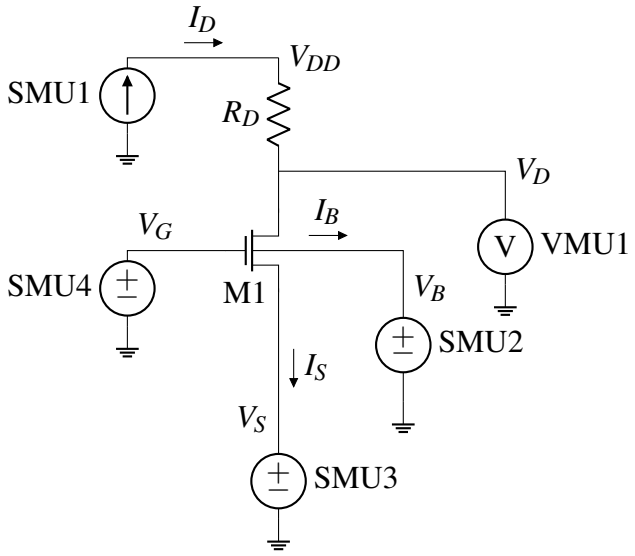


Figure 3.2: Circuit schematic of the test bench for our Soft Error characteristic I-V measurement. The sources (SMU1, 2, 3 and 4) and voltmeter (VMU1) are test units in 4155A. M1 is the N-MOSFET to be tested and modeled (target device).

Table 3.1: Soft Error Test Bench Configuration Details

Source/Monitor Unit	Symbol	Description
SMU1	I_D, V_{DD}	I-Source, Primary Sweep
SMU2	I_B, V_B	Constant-V (-20 V)
SMU3	I_S, V_S	Constant-V (-20 V)
SMU4	V_G, I_G	V-Source, Secondary Sweep
VSU1	V_{SS}	PMOS body pull-down (-20 V)
VMU1	V_D	V-Meter

The test setup for Phase 3 is depicted by Figure 3.2, and the detailed test configuration is listed in Table 3.1. SMU1 drives the drain current I_D by generating a voltage at V_{DD}, while VMU1 measures the actual drain voltage V_D. The resistor R_D introduces an additional node into the circuit (or effectively, an additional degree of freedom). It improves the accuracy of the measurement result by assisting the convergence of the current sweep.

Its value is empirically adjusted and varies between 1 kΩ to 2 kΩ until the Analyzer reports a minimum amount of spurious measurements in the "Snapback" region[9].

Furthermore, all other terminal voltages are driven by separate SMUs, and all terminal currents are monitored. Because of impact ionization-induced body current, $I_B > 0$, and $I_S < I_D$. These three currents I_D, I_S, and I_B are all needed for the Soft Error model extraction, which will be described in Section 3.3.

Meanwhile, the gate current I_G is also measured. Its value should be minimal since the gate dielectric is an insulator. In reality, it should be much less than 1 nA or the measurement error; typically, it is 1–10 pA. If I_G is higher than a threshold such as 1 μA, then we will realize that the gate stack is damaged, and the entire dataset becomes unused. However, we have never experienced such events during actual tests.

The test data is automatically collected and stored on the computer running our program. There are three datasets: low-field, low current (V_{GS}, I_D, G_m) from Phase 1 test; moderate-field, moderate-current (V_{GS}, V_{DS}, I_D) from Phase 2 test; and high-field and/or high-current $(V_{GS}, V_{DS}, I_D, I_S, I_B)$ from Phase 3 test.

We tested twenty-five (25) devices, and each device is tested five (5) separate times, totaling 125 samples each sweep point (a unique combination of voltage-current measurements). Averages are taken across all samples (sample means) with respect to each sweep point, excluding apparently erroneous values due to unexpected measurement failures. Figure 3.3 shows the averaged measurement result of the swept drain current I_D plotted against the monitored drain-source voltage V_{DS}, with branches of different gate-source bias voltages $2.5\,\mathrm{V} \leq V_{GS} \leq 4.5\,\mathrm{V}$.

[9]The term "Snapback region" refers to the measurement data shown in Figure 3.3, which will be shown shortly.

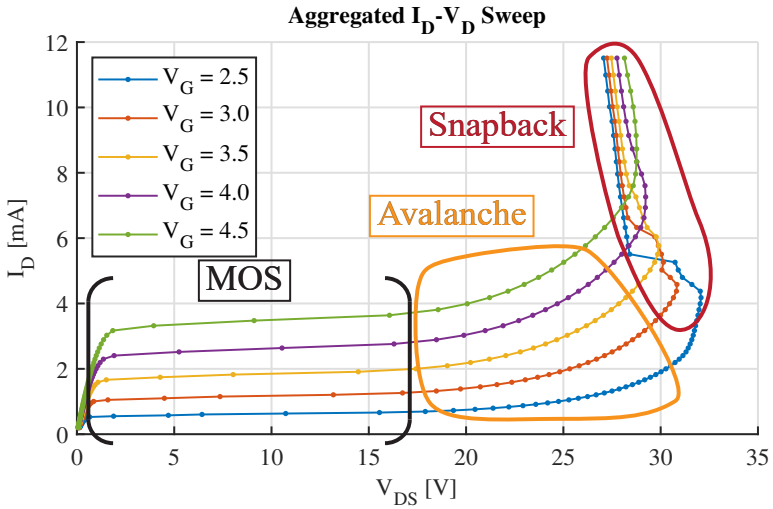


Figure 3.3: I-V curve of the target N-MOSFET device (CD4007) from our experimental measurements. The geometric device parameters extracted using the measurement data are $W = 298\,\mu\text{m}$, $L = 5.87\,\mu\text{m}$, and $t_{OX} = 63.7\,\text{nm}$. The shown data is the sample mean (average) of an aggregated data set using 25 individual N-MOSFET devices, each tested for 5 times. Annotations of different MOSFET operation regions are described in the text.

Annotated in Figure 3.3 are different regions of the MOSFET behaviors. "MOS" includes the regular, textbook-like linear/triode and saturation regions. "Avalanche" is the area of exponential drain-body (impact ionization) and drain-source (BJT) current growth on top of the almost constant saturation current, when channel impact ionization becomes significant, but not enough to cause the drastic reduction in the channel resistance. Finally in the "Snapback" region, the secondary channel current due to the activated parasitic BJT gradually dominates, lateral field collapses, terminal voltage drops with increasing current, and channel resistance effectively reduces to a significantly lower value.

While I_D and V_{DS} are intentionally driven to out-of-design values that are orders of magnitudes higher than the reported typical operating conditions, the gate voltage V_{GS} is held within the operable range described by the device datasheet. Multiple repeated experiments performed on the same device confirm that during the $\sim$ 1-hour test of each device (including 5 consecutive, repeated runs of low-voltage I_D-V_G and I_D-V_D, and high-voltage I_D-V_D tests), there have been no abrupt permanent damage events to the device which could cause a total device failure. The measurement samples of the same sweep point

and from repeated runs with the same device vary only by a minimal amount. Across all low-voltage test (I_D) samples, $\geq 90\%$ of all sweep points (V_{DS}, V_{GS}) have $\leq 30\,\mu A$ of sample stardard deviation ($\sigma(I_D)$); $\geq 93\%$ of all sweep points have $\leq 100\,\mu A$ of maximum sample-average deviation ($\max|I_D - \langle I_D \rangle|$). From the high-voltage test (V_{DS}) samples, $\geq 70\%$ of all sweep points (I_D, V_{GS}) have $\leq 80\,mV$ of sample stardard deviation ($\sigma(V_{DS})$); $\geq 70\%$ of all sweep points have $\leq 320\,mV$ of maximum sample-average deviation ($\max|V_{DS} - \langle V_{DS} \rangle|$).

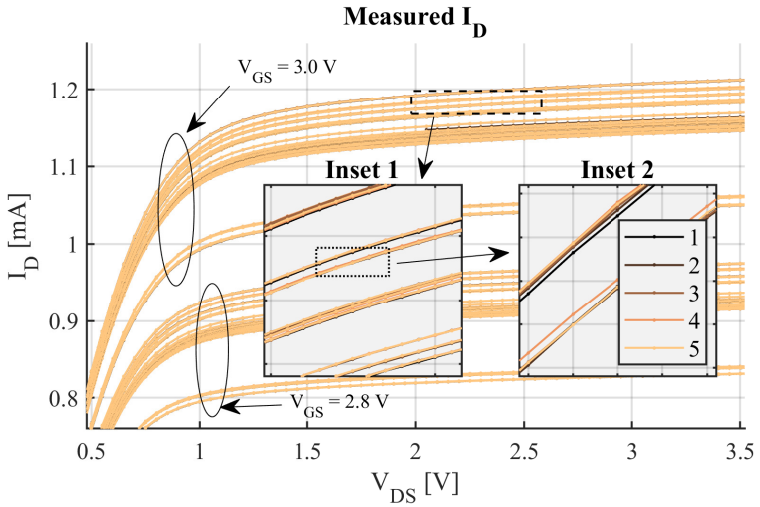


Figure 3.4: Individual I-V measurements of the target N-MOSFET device. For each chip, the repeated measurements (samples) are colored in the order of repetition. Inset 1 shows the data within the dashed-line box. Inset 2 shows that in the dotted-line box in Inset 1. For the same chip, most of the repeated tests yielded nearly identical results, so the curves with different colors are very close to each other. However, the deviation across different chips is much larger.

As shown in Figure 3.4, a low-voltage I_D-V_D sweep is performed prior to each of the 5 repeated high-voltage tests for each of the 25 target devices. The data points are colored with respect to their iteration order. The difference between individual chips is much larger than repeated samples for every single chip. We analyzed the data again after grouping samples from different chips accordingly. In the low-voltage I_D-V_D tests, for 23 of the total 25 tested chips, $\geq 90\%$ of the sweep points (V_{DS}, V_{GS}) have $\leq 0.155\,\mu A$ of sample stardard deviation ($\sigma(I_D)$); for 21 of all chips, $\geq 90\%$ of the sweep points have $\leq 50\,\mu A$ of maximum sample-average deviation ($\max|I_D - \langle I_D \rangle|$). In the high-voltage I_D-V_D tests, the

sweep points are (I_D, V_{GS}) groups, and the output samples are V_{DS}. All tested chips have $\geq 70\%$ of sweep points with $\leq 32\,\text{mV}$ of sample stardard deviation ($\sigma(V_{DS})$); $\geq 70\%$ of all sweep points have $\leq 260\,\text{mV}$ of maximum sample-average deviation ($\max|V_{DS} - \langle V_{DS}\rangle|$). The source/monitor units from the Analyzer generally have $\leq 1\,\text{mV}$ and $\leq 10\,\mu\text{A}$ of measurement error under our particular configuration. Although it is possible that the tests which stress the drain terminal and gate oxide could cause long-term permanent damage to the target devices, according to the observed data, the resulting deviation in device behavior leads us to believe that it is not a major cause of short-term failures related to impact ionization and Snapback.

3.2 Device-Level (DC) Simulation of the Tested Device

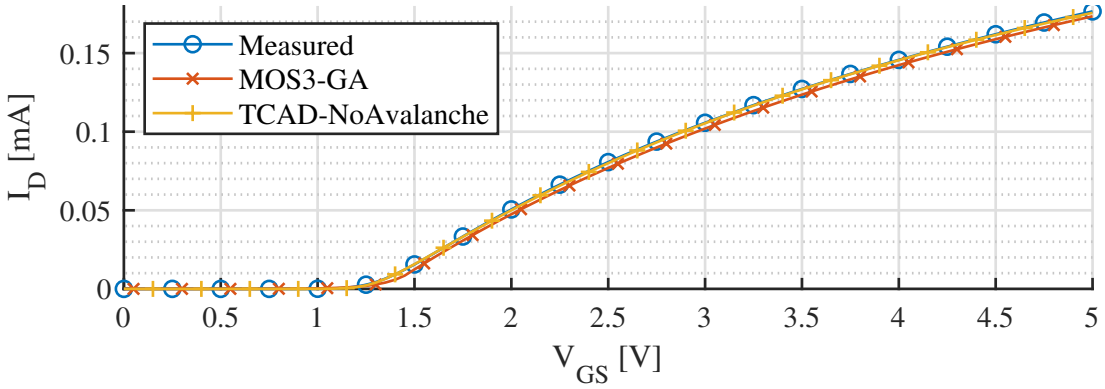
The target N-MOSFET device is simulated to recreate the basic (textbook-like MOSFET channel current) and Snapback (excessive channel current due to impact ionization, parasitic BJT current) behaviors. Since the target device's physical structures and simulation model are unknown to us, we first use an in-house MATLAB program based on the Genetic Algorithm to extract a SPICE MOSFET Level-3 ("MOS3") model [72] from our I_D-V_G and I_D-V_D measurement data. The Genetic Algorithm (GA) implementation will be discussed later in Section 3.3. Then, using the TCAD simulator Cider included in ngspice [62], we create a planar, distributed numerical device model using the extracted geometrical setup and process configuration. Included mobility models are velocity saturation (lateral field), vertical and transverse surface field, and ionized impurity.

Table 3.2: Parameter extraction steps for our proposed Soft Error MOSFET model. The terms "MOS", "Avalanche", and "Snapback" refer to different regions in the I-V data in Figure 3.3.

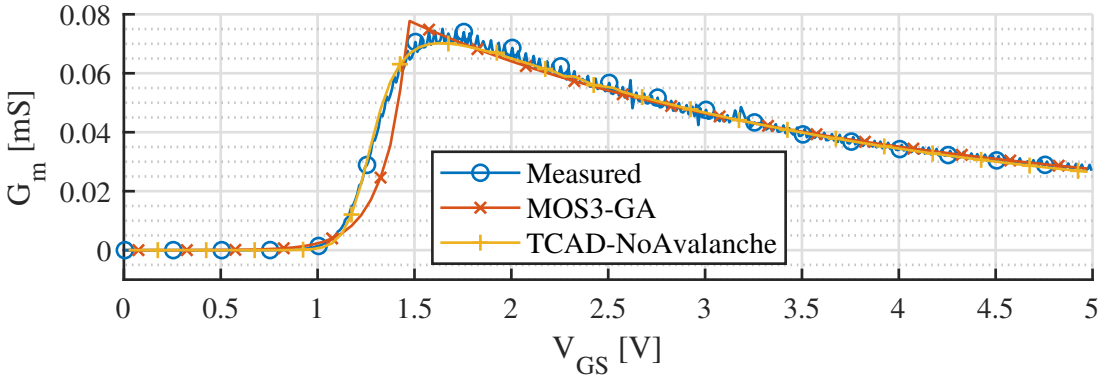
Extraction Steps	Dataset Used	Model Used	Parameters Extracted
Phase 1: Basic MOSFET (low-field)	(I_D, G_m)-V_{GS} ("MOS")	SPICE MOSFET Level-3	$\frac{W}{L}$, TOX, NFS, NSUB, PHI, XJ, THETA
Phase 2: Basic MOSFET (moderate-field)	I_D-(V_{DS}, V_{GS}) ("MOS")	SPICE MOSFET Level-3	W, L, VMAX, DELTA, KAPPA
Phase 3: Snapback (high-field & high-current)	(I_D, I_S, I_B)-(V_{DS}, V_{GS}) ("Avalanche" & "Snapback")	Level-3 and proposed Soft Error model	(listed in Table 3.3)

The textbook-like I-V curves from our measurements (sample average, labeled as "Measured"), the extracted compact model ("MOS3-GA") and the created device-level model without impact ionization calculation ("TCAD-Default") are shown in Figure 3.5. In this step, the "basic" device parameters such as geometrical dimensions, doping concentration, and field-mobility relations are calibrated with our measurement data. The ex-

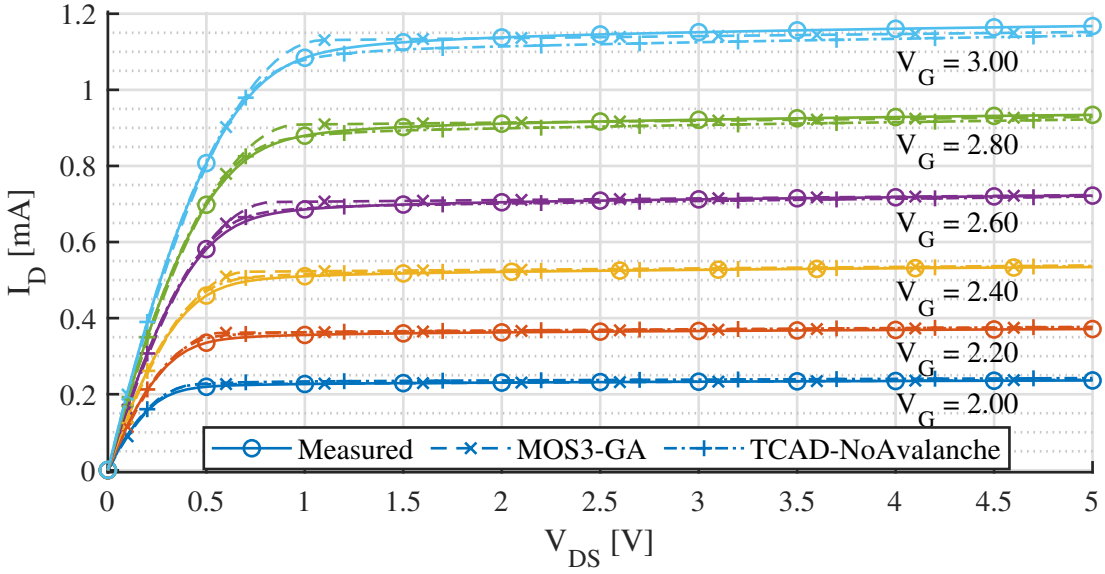
tracted threshold voltage (using the "ELR" method or "extrapolation in linear region" [78]) is $V_{TH} = 1.31\,\mathrm{V}$. Observed from the G_m-V_G curves (Figure 3.5b), the MOS3 model has a delayed turn-on because the first-order discontinuity between the subthreshold and linear regions affects when G_m peaks. But this would not fundamentally change the Soft Error behaviors, since they happen in the saturation region.



(a) Standard I_D-V_G sweep, $V_D = 0.05\,\mathrm{V}$



(b) Transconductance G_m-V_G, $V_D = 0.05\,\mathrm{V}$



(c) Standard I_D-V_D sweep

Figure 3.5: The standard I_D-V_G and I_D-V_D data from our measurement, from the compact model we extracted, and from our numerical model we recreated, labeled with "Measured", "MOS3-GA", and "TCAD-NoAvalanche" (without impact ionization or avalanche calculation), respectively.

It is worth reminding the reader here about our motives. We intend to build a model for the Snapback behavior in a MOSFET, and the model needs to be compatible with practical circuit simulations, so we can evaluate the Soft Error vulnerability when the circuit is exposed to EMI. As explained in Section 2.3, we need a compact model for our purposes. A compact model is a circuit model that represents one or more lumped circuit components using algebraic (closed-form) equations. In this section and the next(Section 3.3), we extract two compact models labeled with "MOS3-GA" and "MOS3+SE-GA".

1. "MOS3-GA" is a SPICE MOSFET Level-3 model extracted with our in-house Genetic Algorithm ("Phase 1" and "Phase 2" in Table 3.2). It represents the regular, textbook behaviors of the tested N-MOSFET. We may consider it as the "baseline" model.

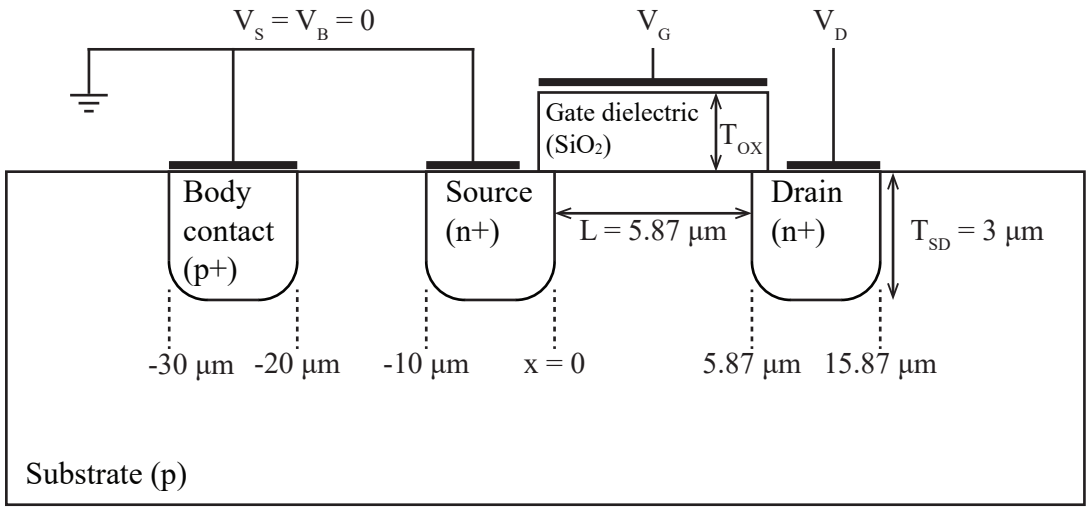
2. "MOS3+SE-GA" contains two models: the "baseline" model from our earlier extraction and the additional Soft Error model representing the Snapback behavior. The second model is also extracted using our Genetic Algorithm ("Phase 3" in Table 3.2).

On the other hand, we perform device-level simulations to verify the underlying physics and help us understand the extracted device structure. For these purposes, we build several TCAD models labeled with "TCAD-NoAvalanche" in Figure 3.5, and three more models labeled with "TCAD-avalanche" in Figure 3.7. We use the information available from our previously extracted "MOS3-GA" model to create a layout structure and use the TCAD application (Cider) to solve the distributed, space-dependent semiconductor equations (2.25). We manually adjust other device-level parameters such as the mobility models and doping profiles to match the test data labeled with "Measured".

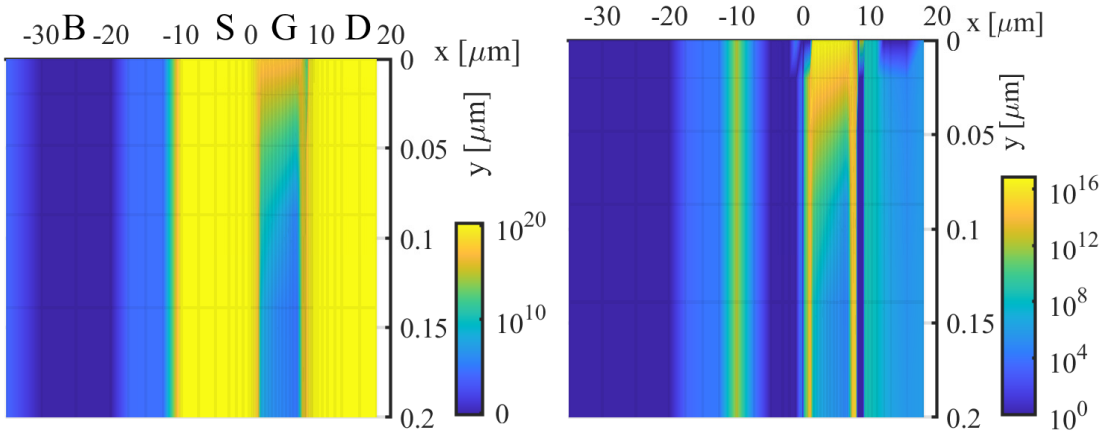
To simulate and observe the impact ionization and Snapback phenomena inside the tested device, the avalanche model is enabled in Cider, which calculates the local generation rate $G_{n,p}$ in Equation 2.20 in Section 2.1. A separate simulation without impact ionization calculation is performed, which provides a "baseline" result for comparison[10].

Shown in Figures 3.6 are some examples of the obtained device-level data with and without impact ionization calculation included.

[10]In reality, impact ionization is always present when the E-field is high enough to produce significant generation rate in Equation 2.20. But we can disable the numerical calculation for comparison.



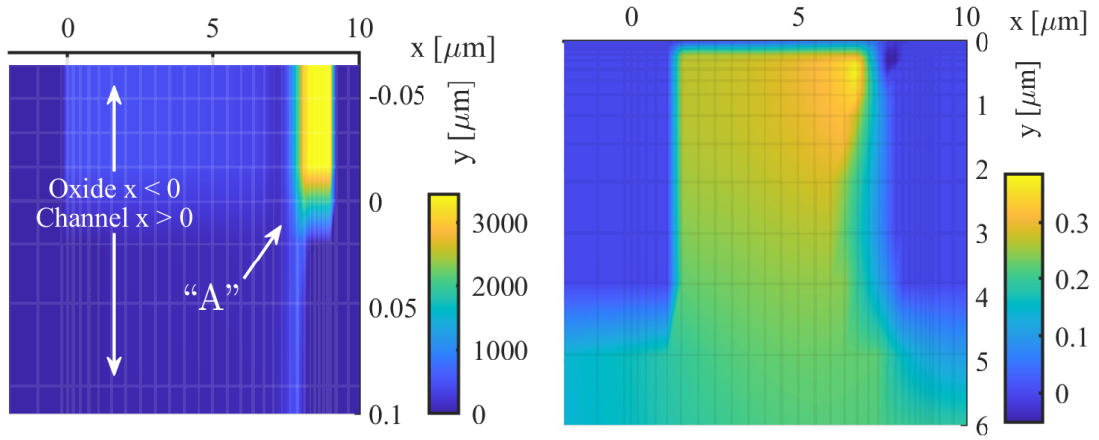
(a) Device geometry



(b) Electron concentration n [cm^{-3}]

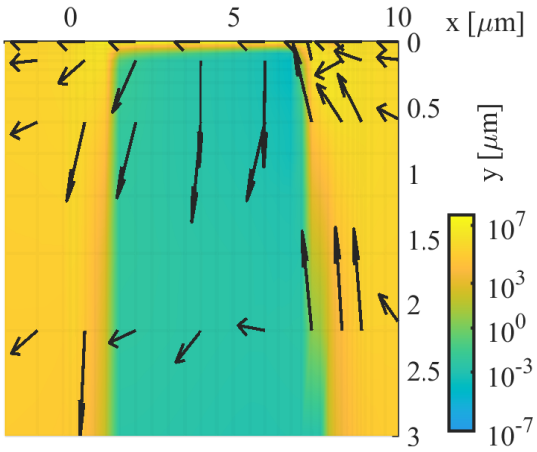
(c) Generated electron concentration Δn [cm^{-3}]

Figure 3.6: Device-level simulation data showing Snapback in a real device which is tested in Section 3.1. $V_{GS} = 3.0\,\text{V}$, $V_{DS} = 25.0\,\text{V}$.
Different x-y space ranges are used for different plots to better expose the helpful data. The geometrical dimensions are $t_{OX} = 64\,\text{nm}$, $W = 298\,\mu\text{m}$, $L = 5.87\,\mu\text{m}$, according to the results from our "Phase 1" parameter extraction described previously. The implant depth T_{SD} for the source, drain, and body contact wells (n++ and p++) are empirically set to $3.0\,\mu\text{m}$ with additional roll-off profiles.
The method "TCAD-avalanche-upgrade-2" from Figure 3.7 is used. The "Δ" values in (c), (e), (g), (i), and (k) are the difference by subtracting the "baseline" results without the avalanche calculation from the results with it.

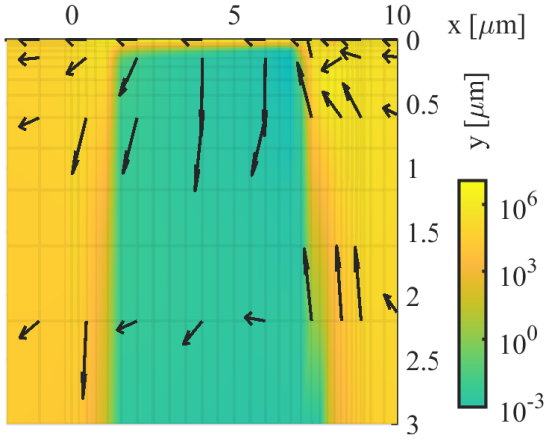


(d) Electric Field
E [kV/cm]

(e) Change in Potential
$\Delta\phi$ [V]

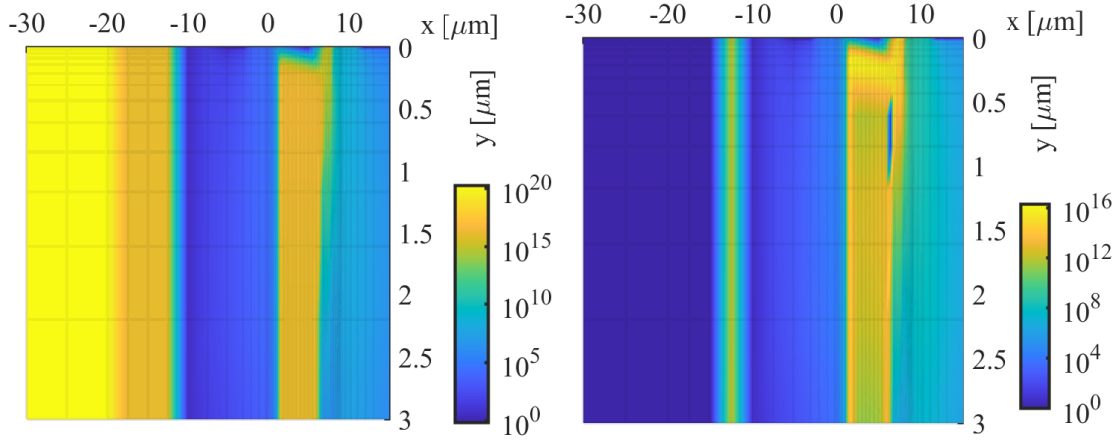


(f) Electron current density
J_n [mA cm^{-2}]



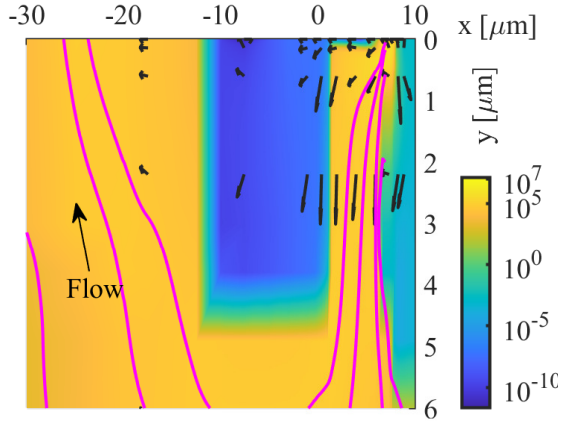
(g) Generated electron current density
ΔJ_n [mA cm^{-2}]

Figure 3.6: (Continued) Device-level simulation data showing Snapback in a real device which is tested in Section 3.1. $V_{GS} = 3.0\,\text{V}$, $V_{DS} = 25.0\,\text{V}$. Different x-y space ranges are used for different plots to better expose the helpful data. More information is given in the previous page.

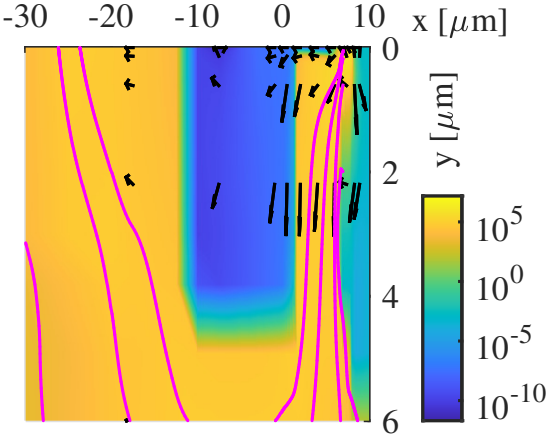


(h) Hole concentration p [cm^{-3}]

(i) Generated hole concentration Δp [cm^{-3}]



(j) Hole Current Density J_p [mA cm^{-2}]



(k) Generated hole Current Density ΔJ_p [mA cm^{-2}] (result without impact ionization calculation is minimal; $\Delta J_p \approx J_p$)

Figure 3.6: (Continued) Device-level simulation data showing Snapback in a real device which is tested in Section 3.1. $V_{GS} = 3.0\,\text{V}$, $V_{DS} = 25.0\,\text{V}$. Different x-y space ranges are used for different plots to better expose the helpful data. More information is given in the previous page.

At high drain terminal bias voltages, because of the much higher lateral field, a significant increase in the electron-hole generation and current density is observed in the channel and substrate. The high field is mostly contained in the vicinity of the drain-body junction (labeled "A" in Figure 3.6d). The peak oxide field is 3.4 MV/cm. The generated electrons ($\Delta n \simeq 1 \times 10^{17}\,\mathrm{cm}^{-3}$ near the interface in Figure 3.6c, by subtracting the "baseline" calculation from the results with avalanche enabled) and holes ($\Delta p \simeq 1 \times 10^{16}\,\mathrm{cm}^{-3}$ at 0.2 µm below the interface in Figure 3.6i) contribute to the additional drain (I_D) and body current (I_B), respectively. As mentioned in Section 2.2.2, the generated holes $\Delta p \simeq 1 \times 10^{15}\,\mathrm{cm}^{-3}$ at $\sim 0.5\,\mu\mathrm{m}$ from the metallurgical junction. The potential increase due to the generated holes is up to $\Delta\phi \simeq 0.26\,\mathrm{V}$ at $\sim 0.5\,\mu\mathrm{m}$ away from the metallurgical junction. The increase in the substrate potential near the body-source junction may activate the parasitic BJT current.

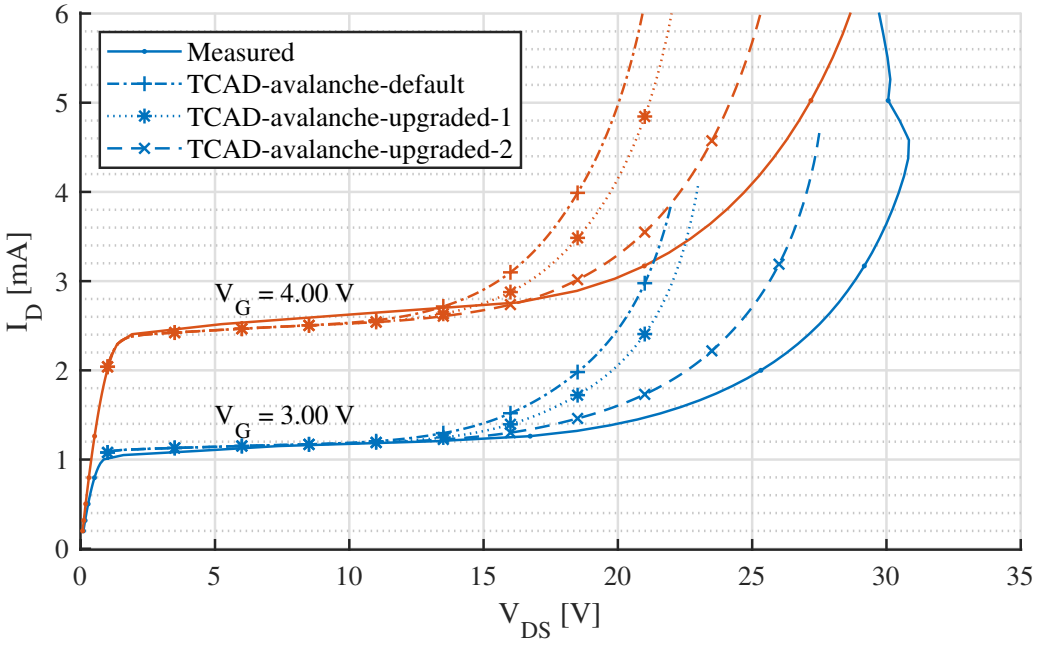


Figure 3.7: Simulated I_D-V_D data with impact ionization included ("TCAD-avalanche-default", "TCAD-avalanche-upgraded-1" and "TCAD-avalanche-upgraded-2") compared to measurement data ("Measured").

In Figure 3.7, multiple sets of data are shown for the drain terminal current-voltage (I_D-V_D) readings. The "Measured" is the test data from the "Phase 3" measurement de-

scribed in Section 3.1 and already shown in Figure 3.3. The "TCAD-avalanche-default" is the terminal readings from device-level simulations using the impact ionization (avalanche) model provided by Cider/ngspice [79]. The "TCAD-avalanche-upgraded-1" and "upgraded-2" are device simulation results using two other methods by modifying the ngspice source code.

It is clear that all simulation results – using the calibrated device structure – are overpredicting the impact ionization current. As a result, the "Snapback point" occurs earlier than measured when increasing the stress current (I_D). Our intent of "upgrading" or modifying the existing model in Cider/ngspice is to improve the simulation by reducing discrepancies.

First, a variation to the original impact ionization model [65] is integrated into the ngspice source code ("TCAD-avalanche-upgraded-1"). The expression and parameters for the local avalanche rate model (Equation 2.19) are calibrated with literature data and different than the "default" model. Next, we change to use only the lateral field and current components to find the ionization rate ("TCAD-avalanche-upgraded-2"). The current components used in Equation 2.21 are both vertical and lateral, meaning that the vertical (gate) field has equal influence in the generation process as the lateral (drain) field. However, we argue that the work done by the vertical field is less because there is less distance for the electron current to flow; as a result, electrons do not get energized in the vertical direction as much as in the lateral direction.

Both results are closer to the measured data; however, there are still discrepancies. To summarize, the device-level simulations explain the origin of Soft Errors, but depending on the computation methodologies, they could be overestimated, and they should be compared against experiments.

3.3 Circuit-Level (DC) Soft Error Model Extraction Results

In this section, we describe our methodology to extract our compact Soft Error model using the MOSFET's characteristic data from our in-house experimental tests.

Ideally, we would want to build a calibrated MOSFET device-level Soft Error simulation model including Snapback, and perform time-domain simulations with practical, functional circuits. However, device-level simulation using the space-dependent drift-diffusion system of equations is generally computationally expensive and time consuming, and thus it is challenging to directly use a device simulator for time-dependent circuit simulations containing more than one MOSFET. Furthermore, the device simulator used in this work experiences convergence issues under high voltages, due to the excessive amount of generated carriers (from impact ionization) which are not included in the self-consistent solving process of the drift-diffusion matrix equation.

Therefore, a compact Soft Error model representing Snapback with lumped equivalent components is necessary for multi-device time-domain simulation. However, it is challenging to extract the parameters for our compact Soft Error model, since the underlying model expressions do not come in ways that are easy to handle mathematically with traditional methods, such as the gradient descent method. The model contains over twenty independent parameters, while three currents (I_D, I_S, I_B) and two voltages (V_{GS}, V_{DS}) are the variables to be matched to experimental data. At the same time, the equivalent circuit including the regular MOSFET, the avalanche current source, and the parasitic BJT and resistor components needs to be solved self-consistently. Moreover, our Soft Error model has gone through several revisions to improve accuracy by including more physics where practical, and each time the equivalent circuit components and the model equations are modified, added, or removed, necessary changes to the extraction process must be taken. Hence, our capability of fast prototyping a compact model greatly depends on the adaptability of our extraction technique. Therefore, we have developed an extraction technique

suitable for our scenario, which is described next. This technique allows for fast prototyping a compact model when the model equations need changing from time to time, and it is compatible with circuit-level SPICE simulations which produce trial results to be programmatically matched to calibration data.

The I-V_{DS} data of the target device are obtained from various methods listed as follows (all methods and results are of our work), and the curves are drawn and compared in Figure 3.8:

- Experimentally measured data (labeled with "Measured")
- Extracted Soft Error and SPICE Level-3 models combined ("MOS3+SE-GA")
- Extracted SPICE MOSFET Level-3 model alone ("MOS3-GA")
- Device-level (TCAD) simulation result with impact ionization calculation ("TCAD-avalanche-upgraded-2")
- TCAD result without impact ionization calculation ("TCAD-NoAvalanche")

For the SPICE Level-3 and the non-ionizing distributed models, it is obvious that none of the three terminal currents exhibits exponential-like growth under high V_{DS}, although the linear-like I_{DS} increase due to channel length modulation can be seen. The combined compact Level-3 and Soft Error model shows the exponential-like current increase and the Snapback behavior. The excessive I_{DS} under lower V_{DS} in this model has been and may be further reduced by fine-tuning the GA extraction search space and improving the model equations.

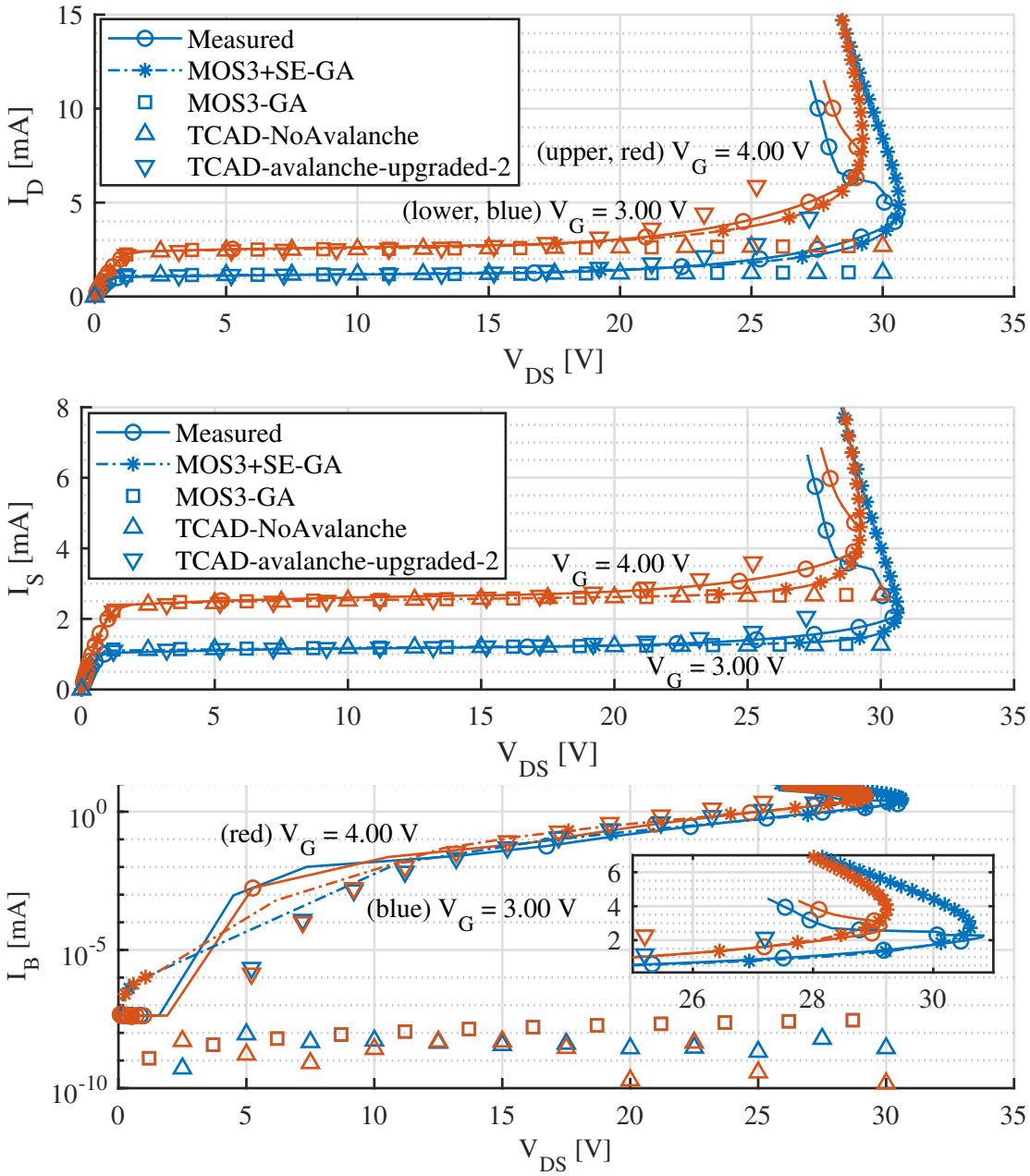


Figure 3.8: Simulated *I-V*$_{DS}$ data from the extracted MOS3 and Soft Error model ("MOS3+SE-GA") of the target device compared to our measurements ("Measured"), extracted SPICE Level-3 model ("MOS3-GA"), distributed model with impact ionization ("TCAD-avalanche-upgraded-2") and without ("TCAD-NoAvalanche") calculations.

3.4 Compact Circuit Model Extraction Technique: An In-House Genetic Algorithm

In this section, we describe our extraction technique used to generate two models used in this work: a SPICE Level-3 MOSFET model for the "basic" MOSFET operations (the "MOS3-GA" model) and a compact or circuit-level Soft Error model representing the Snapback phenomenon (the "MOS3+SE-GA" model).

The extraction method is based on the *Genetic Algorithm* (GA), which partially randomly searches within the entire parameter space given with reasonable boundaries (parameter ranges), keeping a memory of "known-good" solutions over iterations. The flowchart of this method adapted to our compact modeling is shown in Figure 3.9. In one iteration or "generation", a population of trial (tentative) parameter sets is generated. Each set or "gene" contains all parameters needed to perform a complete circuit simulation. The genes are transferred to the circuit simulator (ngspice) along with pre-defined voltage inputs. A cost function or "fitness" is evaluated using the simulated current outputs and calibration data (e.g. from experimental tests), which in general tells the discrepancy between the trial simulation result and given data. At the beginning of the next generation, a new population of genes is generated based on fitness values, favoring those parameters which fit the data better. The iteration is repeated until the fitness reaches an arbitrarily set threshold.

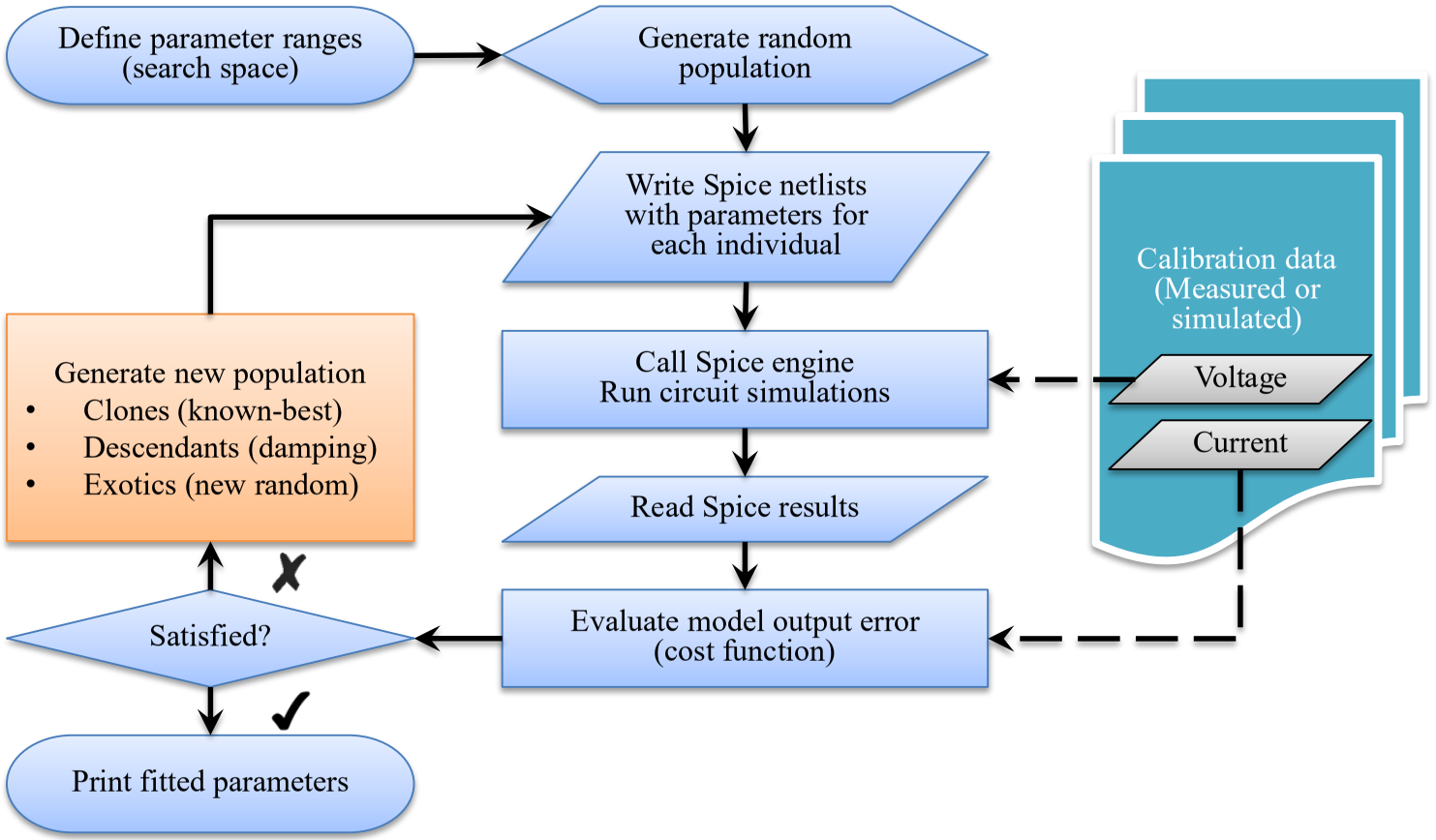


Figure 3.9: Flowchart showing the compact model extraction steps by using the Genetic Algorithm

Several methods are used to generate new parameters, including copying an existing combination, mixing two existing combinations by random ratios, and creating a new combination totally randomly. Generally speaking, a parameter set that produces less discrepancy between the trial simulation result and the supplied data will have a higher "fitness" and a better chance to be preserved. A small collection of "genes" with the best "fitness" will be used to produce more "descendants" than others. In this way, the known-good solutions are "amplified" over iterations. On the other hand, the randomness introduced by the mixing and random terms keeps the search process from getting stuck at a local minimum.

In summary, the Genetic Algorithm is a search method that indirectly follows the trajectory of the system and has Brownian motions as well. Genetic Algorithm is a generic and adaptive technique; it has been previously implemented for different tasks in the microelectronics area [65, 67, 80]. With a good choice of the simulation model and decent fine-tuning of the details, it can become very effective in terms of producing the best fitting parameters.

We implemented the above extraction process in-house. A MATLAB program generates new parameters and evaluates the cost function, which is a weighted RMS error between the data and trial results as follows

$$\text{Fitness} = \sqrt{\frac{1}{N_{\text{data}}} \sum_{\substack{V_{GS} \\ V_{DS}}} \left\{ \left[I_{\text{data}}\left(V_{GS}, V_{DS}\right) - I_{\text{trial}}\left(V_{GS}, V_{DS}\right) \right]^2 \left[\text{Weight}\left(V_{GS}, V_{DS}\right) \right] \right\}} \qquad (3.2)$$

The calibration data $I_{\text{data}}\left(V_{GS}, V_{DS}\right)$ can be obtained from either experimental tests or device (TCAD) simulations. The trial data I_{trial} is the simulation result with one set of tentative parameters (one gene). In this work, we use our experimentally measured I-V characteristics as shown in Figure 3.3 (described in Section 3.1). For the Soft Error model extraction ("Phase 3" in Table 3.2), the Snapback measurement data is used, and I_{data} contains $I_D\left(V_{GS}, V_{DS}\right)$, $I_B\left(V_{GS}, V_{DS}\right)$, and $I_S\left(V_{GS}, V_{DS}\right)$. For the basic MOSFET model extraction (SPICE Level-3 in our study, "Phase 1" and "Phase 2"), the low-voltage $I_D - V_{GS}$, $G_m - V_{GS}$ (described below) and $I_D - V_{DS}$ data is used.

The weighting function can be as simple as a constant value for all data points (uniform weighting) as follows:

$$\text{Uniform weighting:} \qquad \text{Weight} = 1 \tag{3.3}$$

However, we discovered that certain irrelevant factors might negatively influence the quality of the extraction results, and thus using the uniform weighting scheme may not be efficient or even effective.

Specifically in the Snapback measurement data, the three terminal currents I_D, I_S, and I_B show vastly different characteristics as V_{DS} and V_{GS} change, and the three current readings may have discrepancies at different orders of magnitude. The actual data values will be shown later in Figure 3.8; for now, it is important to address the conceptual idea of paying attention to the extraction process. The overall fitness may only reflect the discrepancies in one or two of the three terminal currents, which is not as desired. Meanwhile, simply using relative errors has its problems, too, as the current readings have wide ranges, and even an acceptable solution may have large relative errors in certain regions with small current readings. In contrast, a large absolute error in the high-voltage, high-current region may not register any significant relative error. Therefore, it is necessary to normalize all data groups (e.g. I_D, I_B, and I_S in Amperes in this case, and later G_m in $\mathrm{A/V}$ or Ω^{-1}) to one group (e.g. the current I_D in Amperes) so they become equally important in terms of the magnitude of their values. We achieve so by applying the following weighting function:

Absolute linear weighting for Snapback extraction (Phase 3):

$$\text{Weight} = \begin{cases} 1 & \text{for } I_D \\ \dfrac{I_{D,\max}}{I_{B,\max}} & \text{for } I_B \\ \dfrac{I_{D,\max}}{I_{S,\max}} & \text{for } I_S \end{cases} \tag{3.4}$$

We have also identified other potential challenges, yielding two more weighting schemes as follows. We need to match the basic MOSFET model (SPICE Level-3 in our

study) to the low-voltage data (I_D-V_{GS} and I_D-V_{DS}). The data has previously been shown in Figure 3.5 in Section 3.2. Discrepancies are unimportant as long as they are insignificant, but they can become problematic if the uniform weighting is applied.

First, in the I_D-V_{GS} data, the sub-threshold voltage region ($V_{GS} < V_{TH}$) may have more absolute and/or relative errors than the linear region (full inversion). However, removing the low-voltage portion from the calibration data was overly simplified when we observed that the result matched the above-threshold region perfectly but generated incorrect results below the threshold since they were not calibrated. Plus, we needed to match the large-signal transconductance G_m found by:

$$G_m = \frac{\partial I_D}{\partial V_{GS}} \tag{3.5}$$

Hence, we empirically developed the following weighting formula that depended on the gate voltage V_{GS}, which gradually increases when the gate voltage V_{GS} increases (and the terminal current I_D increases accordingly) and also normalizes the magnitude of transconductance data G_m to match the terminal current:

Voltage-dependent weighting for $I_D - V_{GS}$ extraction (Phase 1):

$$\text{Weight}\,(V_{GS}) = w_1\ w_2 \tag{3.6a}$$

$$w_1 = \left[0.4 + 0.6 \times \frac{V_{GS} - V_{GS,\min}}{V_{GS,\max}}\right]^{0.6} \tag{3.6b}$$

$$w_2 = \begin{cases} 1 & \text{for } I_D \\ \dfrac{I_{D,\max}}{G_{m,\max}} & \text{for } G_m \end{cases} \tag{3.6c}$$

where the units of the quantites are coerced to current (A).

A similar strategy was applied for the I_D-V_{DS} extraction using the basic MOSFET model. In addition to Equation 3.6, the subthreshold and linear regions were skipped because the parameters related to V_{TH} have already been extracted from the I_D-V_{GS} data; changing V_{TH} again could adversely affect the saturation current, which should be modulated by parameters related to channel length modulation. By applying the following

weighting formula, the low-V_{GS} region was excluded, and the data had more "importance" in higher V_{GS} and V_{DS} regions.

Voltage-dependent weighting for $I_D - V_{DS}$ extraction (Phase 2):

$$\text{Weight}\,(V_{GS}, V_{DS}) = w_3\ w_4\ w_5 \tag{3.7a}$$

$$w_3 = \left[0.4 + 0.6 \times \frac{V_{GS} - V_{GS,\min}}{V_{GS,\max}}\right]^2 \tag{3.7b}$$

$$w_4 = \left[0.4 + 0.6 \times \frac{V_{DS} - V_{DS,\min}}{V_{DS,\max}}\right]^2 \tag{3.7c}$$

$$w_5 = \begin{cases} 1 & V_{GS} \geq 1\,\text{V} \\ 0 & \text{otherwise} \end{cases} \tag{3.7d}$$

We developed a Python program that maintains a pool of workers that translate the parameter set, trial simulation input, and output data between the Genetic Algorithm program (MATLAB) and the simulator (ngspice). Performance-wise, this method is strongly prohibitive without enough computational power. A typical extraction converges to a reasonably accurate result after at least $\sim 10^6$ individual trial simulations (entire sets of I-V sweeps), given that the search space (ranges of the fitting parameters) is defined with the knowledge from previous extraction attempts. Each trial takes a few seconds including running the SPICE simulation and all communication overhead time, summing up to a considerably large amount of computation demand.

Therefore, we implemented several network-based multi-processing modules in the Python program so the extraction work can be shared by many PC machines that are connected to the same local-area network. A block-level design schematic is shown in Figure 3.10. Again, the Genetic Algorithm aspect of the extraction is handled by a MATLAB program, also developed by us in-house. Despite the overhead including necessary file exchange, MATLAB function evaluation and parameter generation procedures that are not parallelized, the extraction time is substantially reduced by simply scaling up the total CPU core count in the multi-computer network.

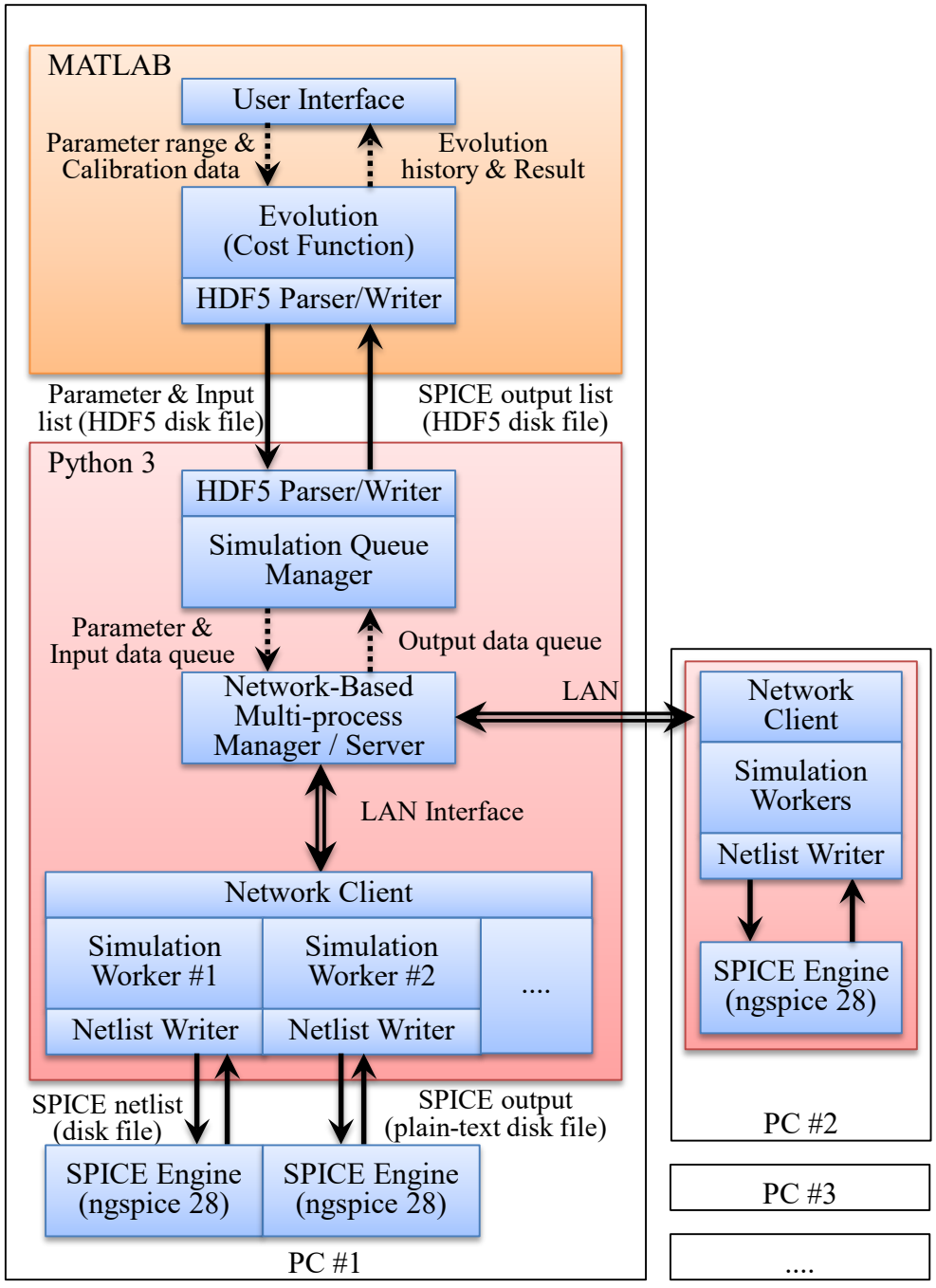


Figure 3.10: Block-level schematic showing the software design to extract the Soft Error model with the Genetic Algorithm

In Figure 3.11, a collection of key results of each iterative generation during the extraction are plotted against the iteration number. The population (number of total trials in one generation) is 2000. The number of generations is 4001, so in total, about 8×10^6 independent SPICE simulations (I_D-V_{DS} sweeps with 128 sweep points) have been executed. For each generation, the best solution which has the lowest "cost" is presented in the plots. Collectively, they represent the evolution history of the extracted parameters. The plots are not labeled with respect to parameter names, although each one has been examined during a post-extraction analysis.

The first graph is the "fitness" or total RMS error calculated by Equation 3.2. Its continuous descent depicts the evolution progress. Generation #0 contains entirely randomly generated parameters within pre-defined search boundaries (the search scope). The RMS error reaches the first "plateau" after the first *few* generations, related to a local minimum. Starting from generation #21, when a "breakthrough" has been found, the best solution descends rapidly to a new local minimum with a much lower error, until it becomes relatively stable again. Similar "breakthroughs" can be seen at generation #47, #108, #1318 and #1910. After around generation #2000, the reduction in total error is relatively insignificant. But from the other graphs, one can observe the further changes in the best solutions, indicating improvements.

The second graph shows the "magnitude" of relative changes in parameters. For example, a value of 10^{-2} means a parameter has increased or decreased by 1 % from its previous value, and 10^{-16} is near the numerical error floor (the "epsilon" in the floating point arithmetic). At the beginning (around the first 100 generations), there was rapid and significant changes among all parameters, while the fitness drastically improved.

Between generations #3500 and #4000, the parameters have changed by up to 5 %, with a few exceptions, although from the previous graph, the total fitting error has not seen much improvement. The change in R_{SUB} or the Ohmic substrate resistance is 12.2 % in generation #3514. Also, between generations #3500 and #4000, it is interesting to notice

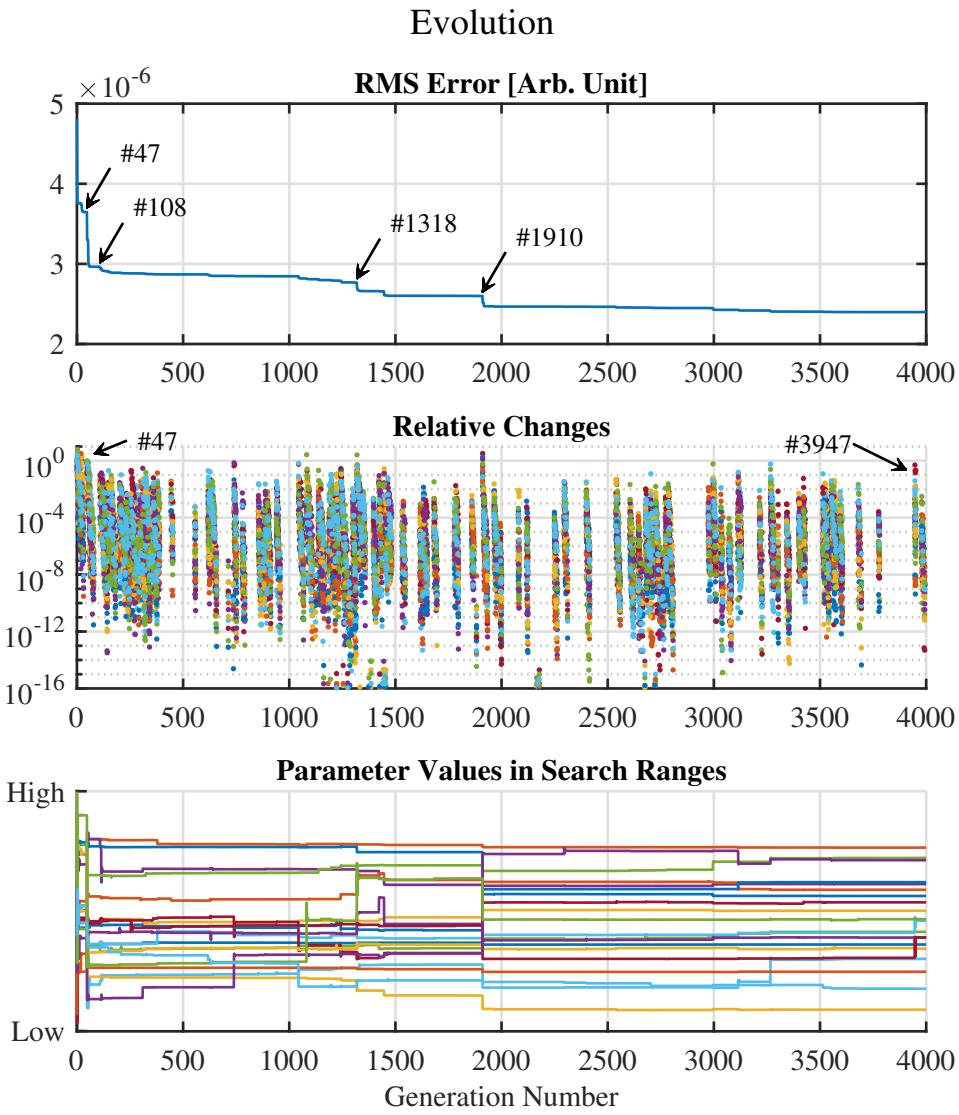


Figure 3.11: Evolution plots of the compact model extraction using the Genetic Algorithm. The data points reflect the best solution of each generation, and are drawn against the number of generations. The plots are, from top to bottom: the cost function value (fitness or total RMS error) of arbitrary unit, with a few generations highlighted by arrows; the magnitudes (always-positive values) of *relative changes* in the individual parameters of the best solution; and the *"relative"* parameter values relative to the pre-defined search ranges. Each line separated by colors refers to one parameter; they are not labeled in this figure for the sake of simplicity.

that the change in β_R is as high as 50.2 % found in generation #3947, knowing that this parameter is actually unused in the actual simulations, but accidentally kept in the model. Meanwhile, another parameter x_1 also changes from 0.331 µm to 0.346 µm. It is used as the lower bound of the drain-body space charge region width in Equation 2.39. Since the drain-body junction is never forward biased in practical circuit applications, this paramater x_1 only serves as a fail-safe mechanism preventing simulation program from crashing (e.g. when user input creates an error, or the SPICE engine attempts a negative V_{DS} during iterations and before reaching convergence). Therefore, generation #3947 has become a demonstration of the evolutional behaviors when a parameter does not change the model's input-output relations.

The third graph shows the actual values of the parameters, scaled to their corresponding pre-defined search ranges denoted by "Low" and "High". From this graph, a *visual check* can be performed to confirm that the parameters stay in local minimums for a certain period of time, rather than constantly increasing or decreasing, which would have been difficult to identify by observing the relative changes. After around generation #3500, the RMS error reaches a relatively stable number, and the relative changes in parameters are generally $< 5\,\%$, despite the visually obvious "jumps" in β_R (which is unused) and in d_1. This is considered as the second phase of the extraction, indicating the extraction has either almost certainly found the global minimum or reached a good approximation. Thus, we believe that the extracted model is good. The extracted Soft Error model parameters are read from generation #4000 and listed in Table 3.3.

Table 3.3: List of extracted Soft Error model parameters for the target device (CD4007)

Symbol	Description	Extracted Value
A	Over-all multiplicative factor (Equation 2.22)	0.230
Impact ionization (lucky-drift) model (Equation 2.40)		
$\mathscr{E}_i^o$	Ionization threshold energy	1.65 eV
F_{S1}	Form shape factor	1.15
F_{S2}	Form shape factor	0.669
λ_R	Average phonon scattering mean free path	11.0 nm
Saturation voltage and depletion region calculation (Equations 2.33 and 2.37-2.39)		
N_{CH}	Channel carrier concentration under strong inversion	1.72×10^{17} cm^{-3}
N_{AS}	Substrate (p-type) doping	3.34×10^{16} cm^{-3}
V_{FB}	Flat-band voltage offset	0.00 V
δv_1	Smoothing factor of saturation voltage	0.0151 V
δv_2	Offset of saturation voltage	-0.0312 V
x_2	Upper limit of drain-body depletion width	5.24 µm
x_1	Lower limit compared to its zero-bias value	0.342
δx_2	Smoothing factor associated to x_2	12.2 nm
δx_1	Smoothing factor associated to x_1	14.7 nm
Substrate resistance (Equation 2.43)		
R_{SUB0}	Body/substrate resistance (Ohmic)	0.763 Ω
a_{di}	Dipole interaction effect, multiplicative factor	0.168
I_{di}	Dipole interaction effect, current offset	4.21 nA
v_{di}	Dipole interaction effect, zero offset	0.0358 V
δv_{di}	Dipole interaction effect, smoothing factor	0.0119 V
Parasitic BJT (Equation 2.42)		
I_{S0}	B-E and C-B junction saturation current	2.15×10^{-17} A
β_F	Forward current gain	2.00×10^3
β_R	Reverse current gain (not used)	3.02×10^3
N_F	B-E junction forward current exponential factor	2.79

Chapter 4: EMI-Induced Soft Error Vulnerabilities: Circuit-Level (Transient) Simulation of Device Vulnerability under EMI Condition

Now that we have developed our Soft Error model in Chapter 2 from physical knowledge and extracted the parameters in Chapter 3 from our experiments, in this chapter, we will simulate a few example circuits in the time domain and at the circuit level to demonstrate the potential application of our compact circuit model for Soft Errors.

Example circuits are first designed using CoolSpice and then simulated iteratively using ngspice in case studies to serve two purposes. First, we demonstrate the proposed Soft Error model's capabilities in transient circuit-level simulations as a design aid to finding any potential signal- and interference-related vulnerabilities in the circuit. We also give examples of the discrepancies between simulation results using the extracted model from Chapter 3.3 and only using the "basic" MOSFET model of the same device.

We simulate EMI conditions as disruptions in the power rail (V_{DD}), which are temporary ("transient") and limited length. It is worth pointing out that by only having disruptions in $V_{DD} = V_{DD}(t)$, our case studies are not exhaustive; the simulated scenarios are less problematic than those when multiple terminals are affected, e.g., V_G and/or V_{SS} (ground line) in addition to V_{DD}. However, our simulations still represent many practical situations since the power rail on a chip is often directly connected to the power plane or power wires on the carrier PCB, making it vulnerable to possible EMI-induced disruptions.

Generally speaking, when the gate terminals also experience voltage disruptions in $V_G(t)$ (mainly via internal connections except at the interfacing front-end stages), the devices are more likely to misbehave or become damaged. It is not our intention to exclude V_{GS} disruptions from the factors leading to Soft Errors; again, our case studies are not

exhaustive. However, by including the calculation of excess drain terminal current and non-linear voltage-current relationship with our circuit model, more vulnerabilities at the drain terminal (I_D-V_D) are exposed to the circuit and system designers. One can always apply different types of EM interference on different functional circuits.

To emphasize, our case studies in this chapter expose potential vulnerabilities due to the impact ionization-induced carrier generation (or avalanche current) and the parasitic BJT structure (or the Snapback phenomenon); these vulnerabilities are mainly triggered by disruptions in the power rail V_{DD}.

Two circuits are simulated, including a resistor-transistor inverter and a two-stage tuned RF amplifier. A summary of the test cases is listed in Table 4.1 with combinations of the test circuit and the type of disruption, as well as the observed changes in behaviors, which will be discussed as follows.

Table 4.1: Summary of tested circuits, types of disruptions and observed behavioral changes

	N-MOSFET Inverter	RF Amplifier
Single Gaussian Unipolar Spikes	Case #1	Case #3
Gaussian-Enveloped Single-Tone Pulses	Case #2	Case #4
Observed disruptions	Additional power consumption; Transient Bit-Flips	Additional power consumption; Decrease in small-signal gain

4.1 Case Study #1: N-MOSFET Inverter under Unipolar Interference on Power Line

The schematic of the simulated circuit is shown in Figure 4.1.

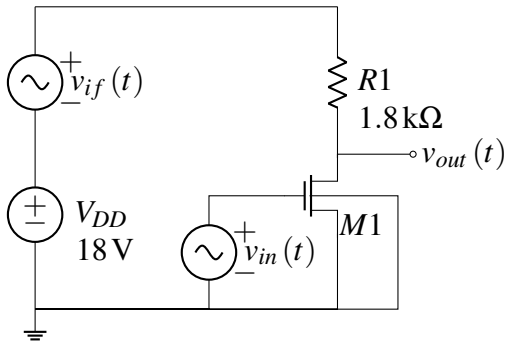


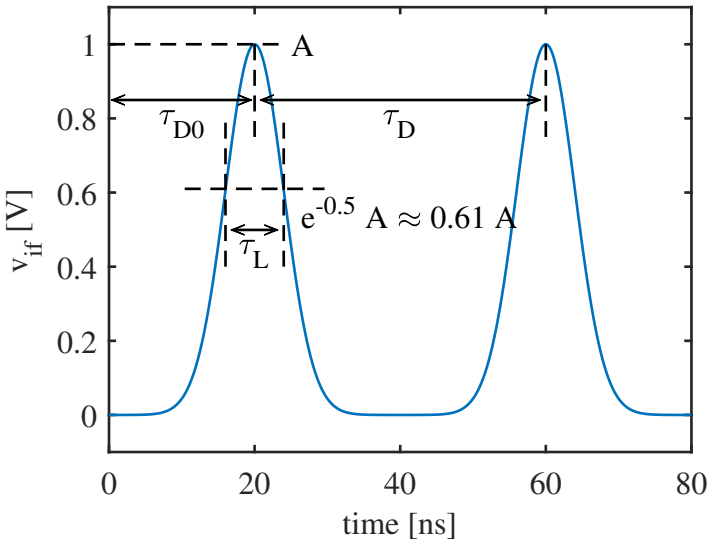
Figure 4.1: Circuit schematic of an N-MOSFET inverter with a disrupted voltage supply.

Transistor M1 is of our primary interest. By changing a single line in the SPICE netlist, we can switch between using the Soft Error model on top of the "basic" MOSFET model and only using the "basic" model for baseline comparison.

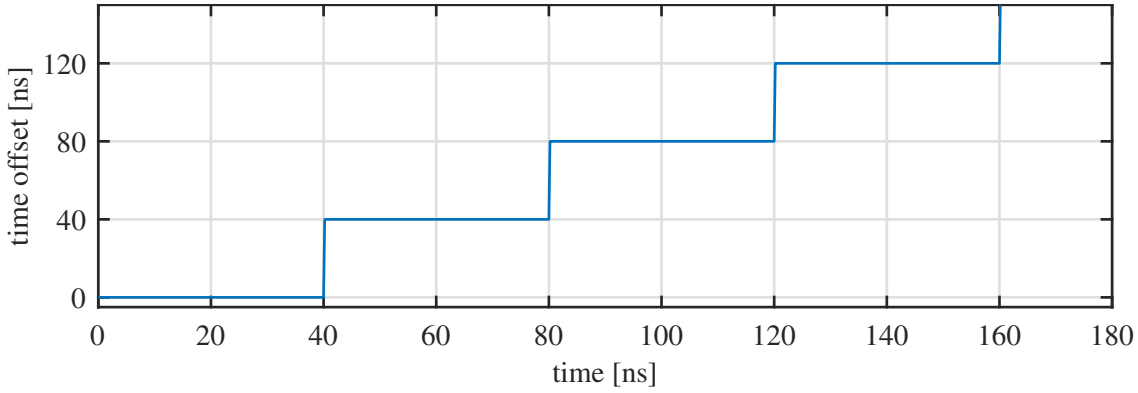
The interference source $v_{if}(t)$ is implemented as a time-dependent voltage source representing a time series of Gaussian pulses. Its analytical expression is defined in Equation 4.1a, and it provides voltage spikes in the V_{DD} rail, as shown in Figure 4.2a, with the unipolar amplitude A, delay τ_{D0}, interval τ_D, and characteristic length τ_L.

$$v_{if}(t) = A\ \exp\left[-\frac{(t-\lfloor t/\tau_D\rfloor\,\tau_D-\tau_{D0})^2}{2\tau_L^2}\right] \tag{4.1a}$$

$$\cong A\ \sum_{n=0}^{N}\exp\left[-\frac{(t-n\tau_D-\tau_{D0})^2}{2\tau_L^2}\right] \stackrel{\text{def}}{=} \widetilde{v_{if}}(t) \tag{4.1b}$$



(a) Time-domain waveform $v_{if}(t)$



(b) Time offset by round-down function $\left(\left\lfloor \frac{t}{\tau_D} \right\rfloor \tau_D\right)$

Figure 4.2: An illustration of the unipolar Gaussian pulse chains $v_{if}(t)$ used in Case #1 and #3. The example time-domain waveform *here* has $A = 1\,\text{V}$, $\tau_D = 40\,\text{ns}$, $\tau_{D0} = 20\,\text{ns}$, and $\tau_L = 4\,\text{ns}$ (which are different than used in Case #1). The round-down function shown in (b) is used to generate $v_{if}(t)$ in Equation 4.1a

The $\lfloor \cdot \rfloor$ is the round-down operator, which returns the nearest integer smaller than the given number. The resulting integer is multiplied by τ_D to produce a stair case-like function value, illustrated in Figure 4.2b, which dynamically changes the Gaussian peak position as time evolves and effectively generates Gaussian pulses repetitively. It is a convenient way to generate time delays of multiples of τ_D in SPICE simulations.

The first definition (Equation 4.1a) is easily implemented in SPICE using a custom

voltage source component (the "B-source"). By using the built-in round-down operator, the Gaussian pulses can be repeated indefinitely, making it convenient to inject the disruptions during different times with respect to the input signal (e.g., rising edge, falling edge, etc.). Besides, we will later use the approximation $\widetilde{v_{if}}(t)$ in Equation 4.1b for frequency-domain analyses since it is easier to apply Fourier transform on the approximated time-domain signal. The approximation holds true as long as the individual pulse peaks are "far apart" ($\tau_D \gg \tau_L$) when the Gaussian tails become negligible before overlapping with other peaks.

Next, we show and analyze the simulation results as below. Two of the MOSFET models are used separately for comparison. The models are introduced and discussed in details in Chapter 3.2, and we repeat the keywords as follows:

- SPICE MOSFET Level-3 model alone ("MOS3-GA"), extracted with our in-house Genetic Algorithm (GA)
- Soft Error and SPICE Level-3 models combined ("MOS3+SE-GA"), extracted with our in-house GA

Shown in Figures 4.3 and 4.4 are the input and output voltage waveforms (MOSFET gate V_{GS} and drain V_{DS}) as well as the power rail $V_{DD}(t)$ within the entire simulation time $t = 1$–$3\,\mu s$. The calibrated models from Chapter 3 are used (the "basic" MOSFET model "MOS3-GA" and the Soft Error model "MOS3+SE-GA"). The input signal $v_{in}(t)$ applied at the MOSFET gate is a 0–20 V, 4 MHz sine wave without disruptions. It is a controlled variable in the case study and does not contain any interference or noise, although applying the Soft Error model does not require so.

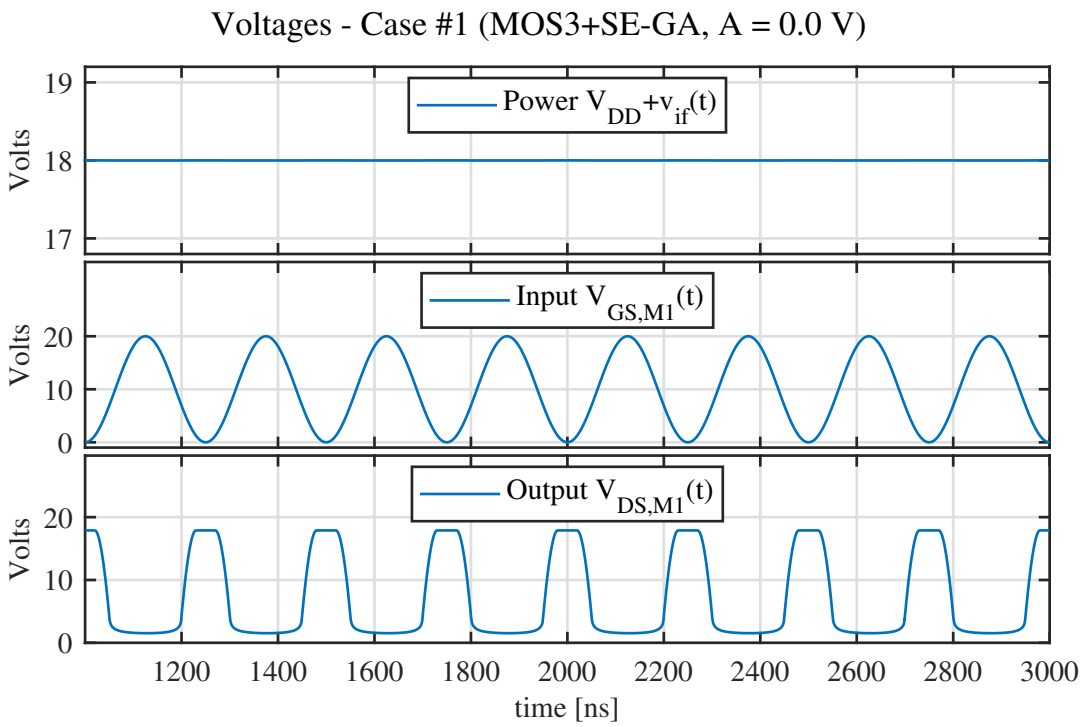


Figure 4.3: Waveforms in Volts of Case #1 in the entire simulation time $t = 1$–$3\,\mu s$. The Soft Error model is enabled ("MOS3+SE-GA"). The input and output are V_{GS} and V_{DS} of M1, respectively. There are no transient disruptions $A = 0.0\,V$.

In Figure 4.3, no disruptions are introduced. The results with and without the Soft Error model are almost identical; for example, the high-level output at V_{DS} is 17.89 V with Soft Error and 17.97 V without it.

In Figure 4.4, the repetitive disruption with amplitude $A = 19.4\,V$ is not enabled until $t = 1\,\mu s$ to ensure the circuit reaches its steady state before becoming affected. The time constants are $\tau_D = 70\,ns$, $\tau_{D0} = 20\,ns$, and $\tau_L = 1\,ns$. Apparently, due to the power voltage spikes, disruptions in the output (drain) voltage are present. The output (V_{DS}) disruptions are more significant when the input is in the low to mid-low range ($V_{GS} \sim 0$–$2\,V$) when the MOSFET channel has barely formed or is not present at all, but the drain voltage is high ($V_{DS} \approx V_{DD} = 18\,V$). In comparison, when the input is near the peak, the drain voltage is rather low ($V_{DS} < 2\,V$) even under disruption, thus less likely to experience significant impact ionization.

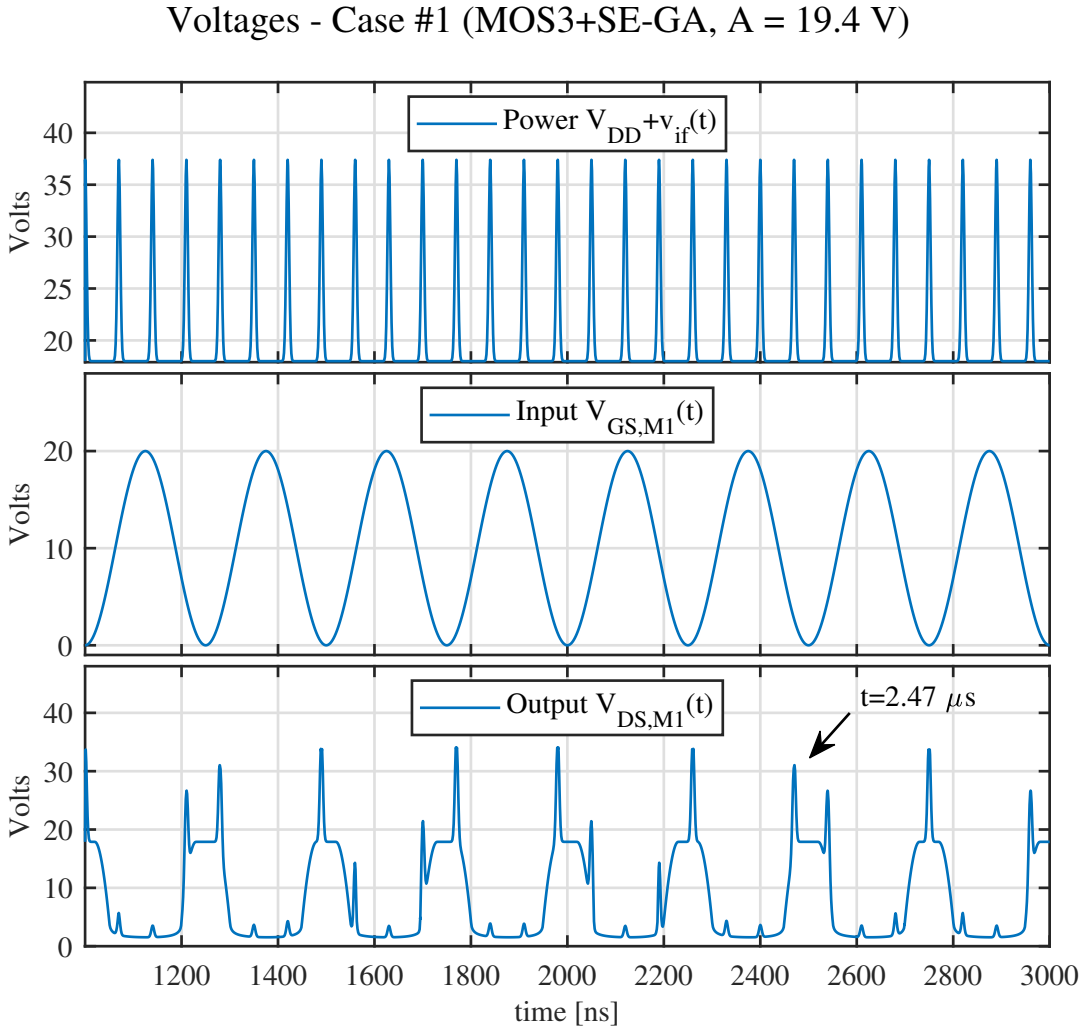


Figure 4.4: Waveforms in Volts of Case #1 in the entire simulation time $t = 1$–$3\,\mu s$. The Soft Error model is enabled ("MOS3+SE-GA"). The input and output are V_{GS} and V_{DS} of M1, respectively. The disruption amplitude is $A = 19.4\,V$. $\tau_D = 70\,ns$, $\tau_{D0} = 20\,ns$, and $\tau_L = 1\,ns$.

On the other hand, in this inverter setup, as V_{DD} increases when it is disrupted, the output voltage V_{DS} increases almost linearly Even without the impact ionization effect, the saturation current increases slightly as V_{DS} increases due to the channel length modulation effect. However, as mentioned above and below, impact ionization and Snapback are causing more problems.

We pick one disruption to provide a more detailed description of the waveform. This is an outstanding example since the "additional" disruptions due to the Snapback phenomenon can be clearly identified by comparing results from with and without using our Soft Error model. In Figure 4.5, the terminal voltages currents of M1 and are shown around time $t = 2.47\,\mu s$. The simulation outputs after applying the Soft Error model ("MOS3+SE-GA") and without it ("MOS3-GA") are both shown.

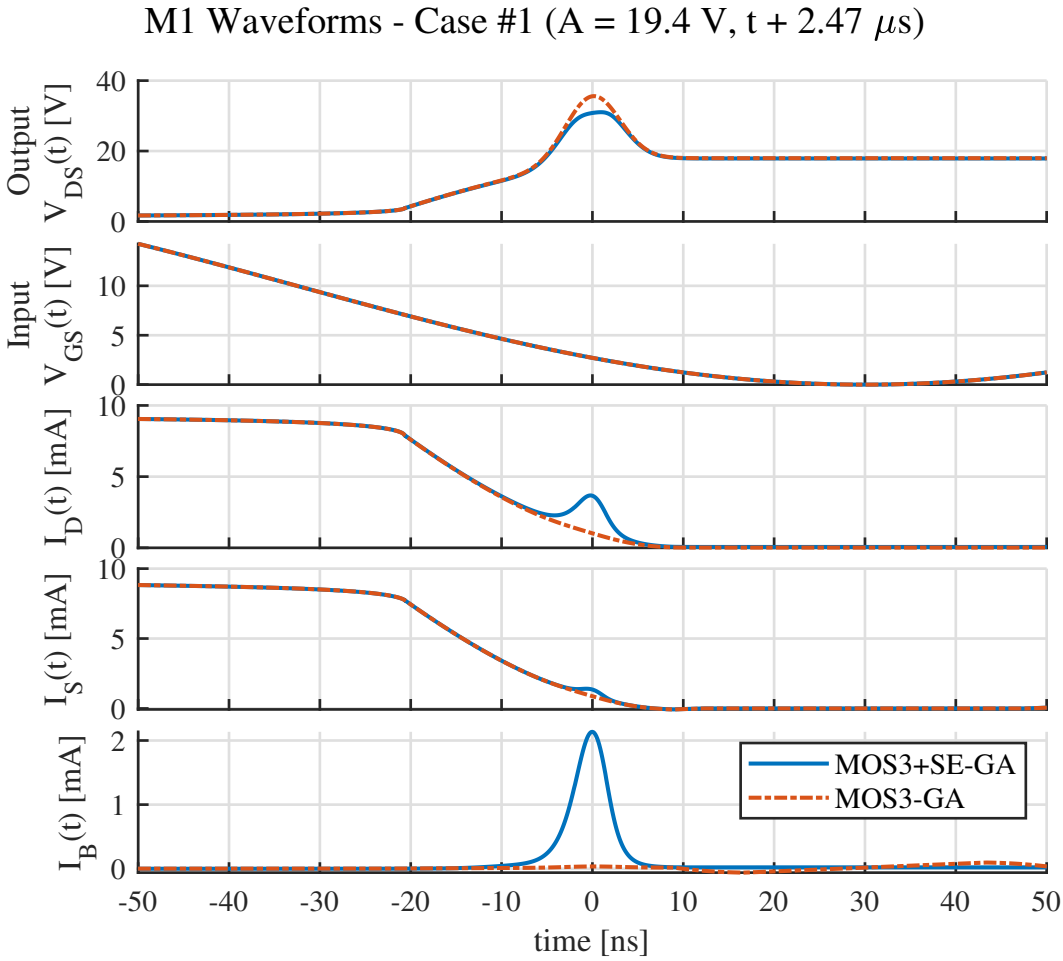


Figure 4.5: Zoom-in waveforms of M1 in Case #1. The time is offset by $(t - 2.47\,\mu s)$. Both results with the Soft Error model ("MOS3+SE-GA") enabled and without ("MOS3-GA") are shown. The disruption amplitude is $A = 19.4\,V$.

By comparing the two results, several changes in the MOSFET behavior are observed after applying the Soft Error model. At time[11] $t = 2.47045\,\mu s$ (or $t = 0.45\,ns$ in Figure 4.5), $V_{GS} = 2.63\,V$. The peak V_{DS} is lowered from 35.5 V to 31.0 V. There is no noticeable elongation of the V_{DS} disruptions, indicating that such disruptions are still "transient". I_D increases from 0.934 mA to 3.46 mA. I_B mainly consists of the impact ionization hole current, and increases from minimal (35.5 µA) to 2.06 mA. The increase in I_S from 0.791 mA to 1.23 mA suggests that the source-body junction is under forward bias, and the parasitic BJT current is present.

Digital bit errors can occur under this type of disruption. We notice that the bit errors are present no matter if our Soft Error model is applied or not, although they are intensified through the Snapback phenomenon. These errors captured by the Soft Error model are potentially problematic in a large-scale circuit.

Besides, the additional terminal currents under excessive voltages can lead to unwanted power consumption. Using the simulation data, we evaluate the instanteneous power $p = vi$ consumed by M1. The average power dissipation is found by

$$\langle P_{M1} \rangle = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} i_{D,M1}(t)\ v_{DS,M1}(t)\, dt \tag{4.2}$$

with $t_2 = 3\,\mu s$, $t_1 = 1\,\mu s$ is drown against various disruption amplitude A added to V_{DD}. The additional power dissipation under disruptions is defined by

$$\Delta \langle P_{M1} \rangle = \langle P_{M1} \rangle - \langle P_{M1,0} \rangle \tag{4.3}$$

where $\langle P_{M1,0} \rangle$ is the power consumption under no disruptions (disruption amplitude $A = 0\,V$), namely the "y-intercept".

[11] Our simulation time step is 0.01 ns.

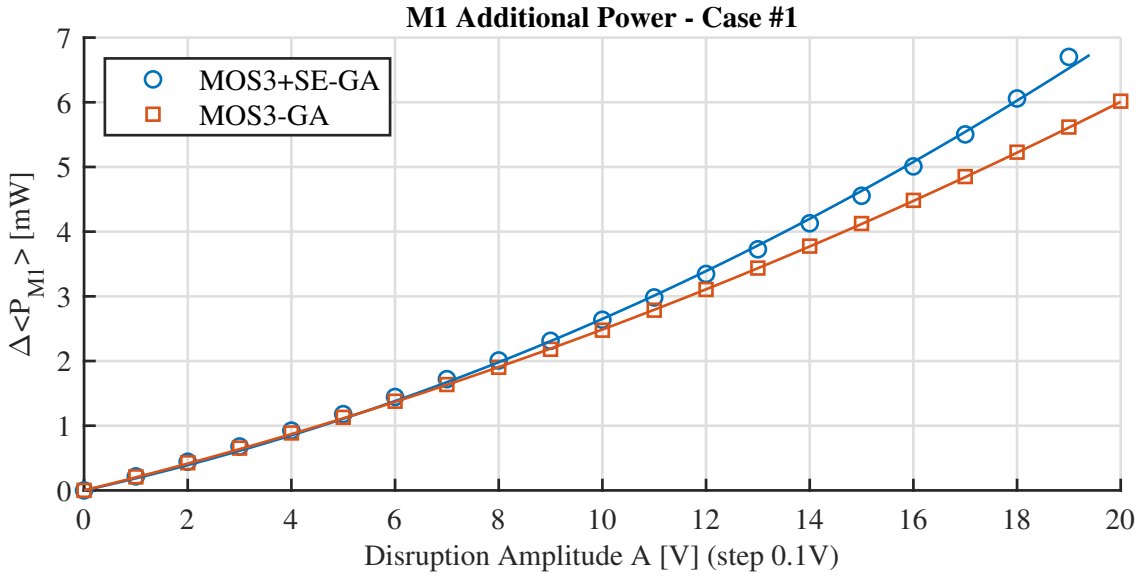


Figure 4.6: Additional power consumed by transistor M1 in Case #1. Results with and without the Soft Error model are labeled as "MOS3+SE-GA" and "MOS3-GA", respectively. The amplitude A (x-axis) in the simulation data is much denser than illustrated with a uniform interval of 0.1 V. The simulation data points (symbols) and fit curves (solid lines) represent the *additional* power $\Delta\langle P_{M1}\rangle$ in Equation 4.3.

In Figure 4.6, $\Delta\langle P_{M1}\rangle$ is drawn against the disruption level in $V_{if}(t)$, quantified by the unipolar pulse amplitude A on top of the DC supply V_{DD}. The plots are offset by their corresponding y-intercept $\langle P_{M1,0}\rangle$. Results after applying the Soft Error model ("MOS3+SE-GA") and without it ("MOS3-GA") are shown. To summarize the "extra" vulnerability due to the Snapback phenomenon, we empirically fit the two sets of data with two quadratic polynomials. With A in V and $\Delta\langle P_{M1}\rangle$ in mW, the fitting formulas are:

with Soft Error model ("MOS3+SE-GA", fit line for the blue circles):

$$\Delta\langle P_{M1}\rangle = 8.71\,A^2 + 0.178\,A \tag{4.4a}$$

without Soft Error model ("MOS3-GA", fit line for the red squares):

$$\Delta\langle P_{M1}\rangle = 8.17\,A^2 + 0.197\,A \tag{4.4b}$$

The power consumption under no disruptions is $\langle P_{M1,0}\rangle = 16.2\,\text{mW}$ with the Soft Error model and 15.9 mW without it. The 2 % difference is due to an insignificant presence of the impact ionization effect and avalanche drain-body current.

Overall, the additional power consumption increases as the disruption level increases. This is expected for the cases with and without our Soft Error model, because the volt-ampere product is generally higher.

In the entire range of disruption amplitude being surveyed, we observe a consistent increase in $\Delta\langle P_{M1}\rangle$ after the Soft Error model is enabled, by up to 21 % at $A = 19.4\,\mathrm{V}$. From the empirically fit quadratic expression, the larger second-order coefficient with the Soft Error model, along with the smaller first-order one, indicates that when the channel impact ionization and parasitic BJT currents are considered, the additional power consumption under high-level disruptions may be much higher than that is predicted by the "basic" MOSFET model. The change depending on disruptions is 68 % and -9.6 % in the quadratic and linear coefficients, respectively; the increase in the no-disruption power is 1.9 %.

Elevated power consumption may become a worrisome problem under certain circumstances, even though the disruption level found in our case study does not seem to cause drastic changes in the circuit behaviors. For example, systems running on batteries may go out of service prior than designed, and the associated additional Joule heating may lead to reliability issues. A circuit or system designer may use the proposed Soft Error model as a tool to test and evaluate a MOSFET circuit's behaviors under EMI exposure. At this point, a link between short-term, transient errors and long-term progressive degradation can be seen; but we still want to focus on the transient aspect of Soft Errors within our scope of study.

Before concluding Case #1, we discuss a detailed aspect as follows. As has been pointed out in Chapter 1, external interferences may cause unwanted resonance in the functional circuit, intensifying the disruption. Through the following calculations, we show that the Gaussian pulse trains used in Cases #1 and #3 have their major frequency components much higher than the operating frequency of the circuits. Thus, they do not introduce major complications due to "resonating" with the circuit or the input signal $V_{GS}(t) = v_{in}(t)$.

We apply Fourier transform to the approximated expression $\widetilde{v_{if}}(t)$ in Equation 4.1b, and knowing that the Fourier transform is an additive operation, the result is:

$$\mathscr{F}\left\{\widetilde{v_{if}}(t)\right\}(f) = \frac{\sqrt{2\pi}A}{f_L} e^{-j2\pi\tau_{D0}f} \sum_{n=0}^{N} \exp\left[-2\pi^2\left(\frac{f}{f_L}\right)^2 - j\frac{2n\pi}{f_D}f\right] \tag{4.5}$$

where $f_D = \tau_D^{-1}$ and $f_L = \tau_L^{-1}$. In Case #1, $f_D = \tau_D^{-1} = (70\,\text{ns})^{-1} = 14\,\text{MHz}$ and $f_L = \tau_L^{-1} = (1\,\text{ns})^{-1} = 1\,\text{GHz}$. They are much higher than the inverter's operation frequency 4 MHz.

The normalized spectrum magnitude as a function of frequency $\left|\mathscr{F}\left\{\widetilde{v_{if}}(t)\right\}\right|(f)$ is shown in Figure 4.7 for various numbers of terms used for the summation.

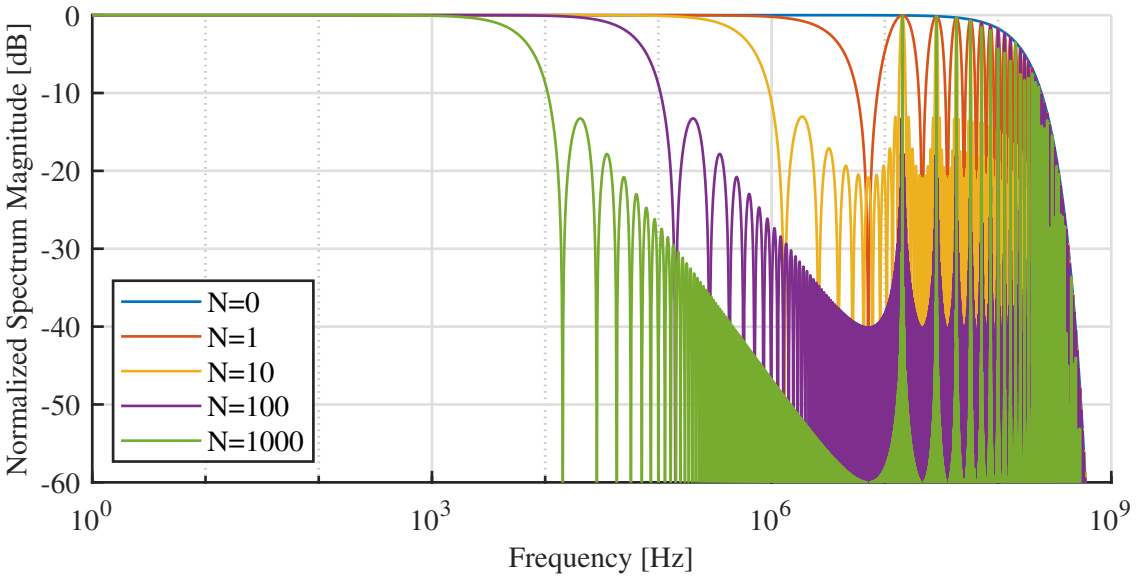


Figure 4.7: Normalized spectrum magnitude for the unipolar Gaussian pulses $\left|\mathscr{F}\left\{\widetilde{v_{if}}(t)\right\}\right|(f)$ used in Case #1 and #3. Various numbers of pulses N are used in the summation in Equation 4.5.

Comments:

1. The Fourier transform of the time-domain function in Equation 4.1b can be easily obtained by applying basic Fourier transform pairs and the convolution theorem, acknowledging that the time-domain function can be treated as the convolution of an infinite train of delta impulse functions and a Gaussian function.

2. When $N = 0$, there is a single Gaussian pulse in the entire time, and its Fourier

transform is also a Gaussian function, with its peak at $f = 0\,\text{Hz}$ (DC) and $-3\,\text{dB}$ point[12] at $f_{-3\,\text{dB}} = f_L\sqrt{\frac{\ln 2}{4\pi^2}} = 133\,\text{MHz}$.

3. When there are more than one pulse in the entire time, i.e., $N \geq 1$, there are infintely many peaks in the spectrum magnitude. For example, when $N = 1$, we can explicitly evaluate the summation as below:

$$\begin{aligned}
\mathscr{F}\left\{\widetilde{v_{if}}(t)\right\}(f) &= \frac{\sqrt{2\pi}A}{f_L}e^{-j2\pi\tau_{D0}f}\sum_{n=0}^{1}\exp\left[-2\pi^2\left(\frac{f}{f_L}\right)^2 - j\frac{2n\pi}{f_D}f\right] \\
&= A'\sum_{n=0}^{1}\exp\left[-j\frac{2n\pi}{f_D}f\right] \qquad (4.6a) \\
\left|\mathscr{F}\left\{\widetilde{v_{if}}(t)\right\}\right|^2(f) &= A'\left[\exp(0)+\exp\left(-j\frac{2\pi}{f_D}f\right)\right]\left[\exp(0)+\exp\left(+j\frac{2\pi}{f_D}f\right)\right] \\
&= 2A'\left(1+\cos\frac{2\pi}{f_D}f\right) \qquad (4.6b)
\end{aligned}$$

where the new coefficient is $A' = \frac{\sqrt{2\pi}A}{f_L}e^{-j2\pi\tau_{D0}f}\exp\left[-2\pi^2\left(\frac{f}{f_L}\right)^2\right]$, and it has three parts: a frequency-independent real number $\frac{\sqrt{2\pi}A}{f_L}$, a frequency-dependent complex-value function $e^{-j2\pi\tau_{D0}f}$ with a constant magnitude of 1, and a frequency-dependent real-value function $\exp\left[-2\pi^2\left(\frac{f}{f_L}\right)^2\right]$ which decreases monotonically as the frequency increases (giving the envelope as seen in Figure 4.7). The cosine function oscillating in the frequency domain gives rise to the spectrum peaks mentioned above; equivalently, they are high-frequency harmonics of the base tone. In Case #1, the base frequency is at $f_D = 14\,\text{MHz}$, and the higher-frequency peaks are at $2f_D = 28\,\text{MHz}$, $3f_D = 42\,\text{MHz}$, and so on.

4. When $N > 1$, $\left|\mathscr{F}\left\{\widetilde{v_{if}}(t)\right\}\right|^2(f) = A'\sum_{n=0}^{N}\exp\left[-j\frac{2n\pi}{f_D}f\right]$. More local peaks in the frequency spectrum are introduced, but those already present when $N = 1$ are still the dominating components; i.e., their amplitudes are far greater. Meanwhile, these "major" components become narrower so more and more "minor" peaks with much lower

[12]The $-3\,\text{dB}$ point is defined as the frequency when the voltage magnitude is $\frac{1}{\sqrt{2}}$ times its peak value. The $-6\,\text{dB}$ point is defined as the frequency when the voltage magnitude is a half of its peak value and is $f_{-6\,\text{dB}} = f_L\sqrt{\frac{\ln 2}{2\pi^2}} = 187\,\text{MHz}$.

amplitudes can be accommodated. The base frequency f_D does not change, nor does the separation between the harmonics, also f_D. This situation is reflected in Figure 4.7, where the normalization process makes all results with different N appear to have the same amplitude. In reality, however, the unipolar pulse chain will introduce a DC offset that depends on the number of pulses over all time.

5. When $N \geq 1$, the entire spectrum is modulated by the Gaussian envelope found when $N = 0$. Since its cut-off frequency is formerly found to be $f_{-3\,\mathrm{dB}} = 133\,\mathrm{MHz} \approx 9 f_D$, before the spectrum becomes effectively negligible, we can observe about a dozen narrow bands. In other words, most of the disruption energy is concentrated in the base frequency f_D and the first several harmonics: 14 MHz, 28 MHz, 42 MHz, and so on. Meanwhile, the functional circuit (the inverter) in this case operates at 4 MHz.

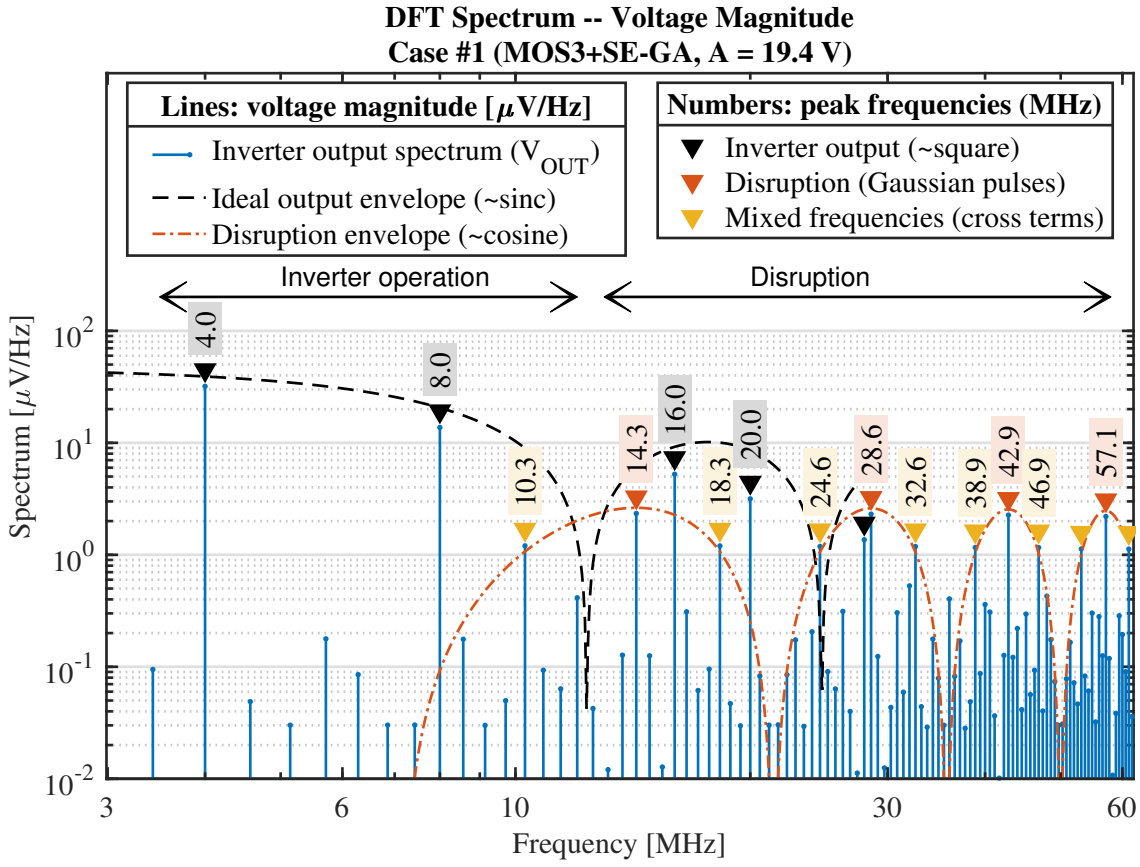


Figure 4.8: Voltage spectrum magnitude of the discrete Fourier transform of the simulated output waveform $|\mathscr{F}\{V_{OUT}(t)\}|(f)$ in Case #1.

At last, we examine how the disruptions affect the output waveform in the frequency domain. We apply a discrete Fourier transform on the time-domain output voltage waveform $V_{OUT}(t)$ and take its magnitude. The resulting spectrum $|V_{OUT}(t)|$ is shown in Figure 4.8. Detailed analyses are illustrated in the graph and listed as below:

1. The time-domain simulation length is 17.5 µs or 70 whole operation cycles (low-high transitions), which is about 8 times longer than displayed in Figure 4.4. This is done in order to increase the frequency-domain resolution to $\Delta f = (17.5\,\mu s)^{-1} \approx 57.1\,kHz$. The sampling frequency is $f_s = 2\,GHz$. The spectrum is normalized[13] to conform power conservation.

2. The inverter's expected output is a rectangular pulse train in the time domain. Ideally, its discrete Fourier transform is a sequence of delta pulses (i.e., matchstick-shaped components) that are non-zero at separated single-frequency locations and have a sinc function[14] envelope. The base-frequency component is found at $f_0 = 4.0\,MHz$, and the two most-significant harmonics are found at 8.0 MHz and 16.0 MHz.

3. The disruptions in the power rail (V_{DD}) lead to additional spikes in the frequency domain. As previously discussed, the disruptions appear in the magnitude spectrum as narrow peaks enveloped by a cosine-like function (Equations 4.5 and 4.6b). The lowest-frequency peak is at $f_D = 14.3\,MHz$.

4. Peaks are also found at the "cross-term" frequencies such as $f_D \pm f_0 = 10.3\,MHz$ and 18.3 MHz, but the are more than 10 times smaller than the inverter's operational output base frequency component at $f_0 = 4\,MHz$. Hence, we do not suspect frequency mixing is a primary concern in this and subsequent case studies.

Therefore, the AC components do not induce resonance since the circuit is not tuned to the particular frequency components in the interference. Instead, the temporary

[13] Theoretically, energy conservation is ensured by Parseval's theorem: $\int_{-\infty}^{+\infty} |v(t)|^2\,dt = 2\pi \int_{-\infty}^{+\infty} |v(f)|^2\,df$. In our case, since we have a finite-length time data (e.g., $t_1 \le t \le t_2$), we interpret Parseval's theorem by *power conservation*: $(t_2 - t_1)^{-1} \int_{t_1}^{t_2} |v(t)|^2\,dt = 2\pi\,(t_2 - t_1)^{-1} \int_{-f_s/2}^{+f_s/2} |v(f)|^2\,df$ where f_s is the sampling frequency of the discrete Fourier transform.

[14] By definition, $\mathrm{sinc}(x) = \frac{\sin \pi x}{\pi x}$.

increase in the power rail $V_{DD} + v_{if}(t)$ is of major concern, as seen in the results above. The origin of the interference could be a temporary error in the power supply rail, and the disrupting voltage (the pulse chain) contains both DC and AC components. When comparing the results from using the "basic" MOSFET model only against using the proposed and calibrated Soft Error model, as shown in Figures 4.5 and 4.6, the differences in output waveforms and power consumption are not directly related to any "resonance" between the function circuit and the disruptions. Instead, it is mainly due to the local generation from impact ionization and the parasitic BJT.

To conclude the Case Study #1, spikes in the output waveforms induced by disruptions may cause "bit flips", which could directly cause a digital circuit without proper error-checking mechanisms to crash. By applying the Soft Error model, the vulnerability situation generally becomes more intense. Besides, no elongated bit errors are observed after the transient disruption disappears. However, the additional power consumption under disruptions suggests a battery-powered circuit may have a shorter operation time than predicted by the "basic" MOSFET model.

4.2 Case Study #2: N-MOSFET Inverter under Symmetric Interference on Power Rail

The functional circuit under test is the same as in Case #1, but the disruptions are now produced by an external radio-frequency interference that couples into the power delivery network on the printed circuit board (PCB). As shown in Figure 4.9, the test circuit in simulation consists of four parts: the un-disrupted DC power supply V_{DD}, the LCR network representing the wire tracing of V_{DD} on the PCB, the time-dependent AC source $i_{if}(t)$ together with an associated inductor L_{if} representing the coupled interference from external EM radiation on the path of V_{DD}, and the functional circuit (N-MOSFET inverter) including R1 and M1. A decoupling capacitor (C_{DD}) is present; however, its effectiveness may depend on its location on the PCB.

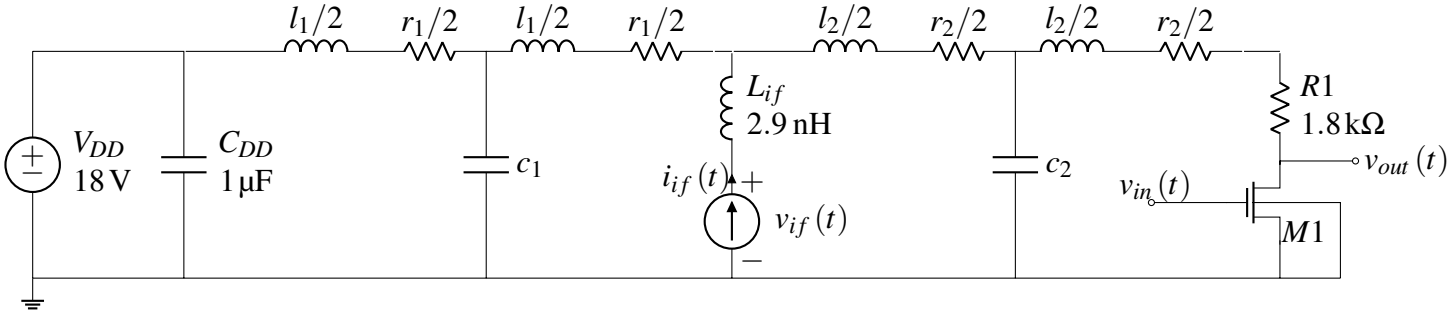


Figure 4.9: The circuit simulated in Case #2, showing the coupling effect of external EMI due to PCB trace structures and the functional circuit under test (N-MOSFET inverter).

The PCB tracing has a total length of 2.00 cm, and the coupling point is in the middle. We choose the following dimensions for the V_{DD} trace, which is a copper wire (electric resistivity $\rho_r = 1.68 \times 10^{-8}\,\Omega\text{m}$, mass density $\rho_m = 8.96\,\text{g/cm}^3$). The width is $W = 2.50\,\text{mm} \approx 100\,\text{mil}$. The thickness is calculated for $1\,\text{oz/ft}^2$ coating, which is $T = 34.0\,\mu\text{m}$. The ground network is considered as an entirely filled plane with much smaller LCR values. The power-ground distance is $H = 1.20\,\text{mm} \approx 50\,\text{mil}$. The dielectric constant is set at $\varepsilon_r = 4.50$ according to typical FR4 material properties.[15]

[15] FR4 is a very widely used material to build printed circuit boards (PCB).

Using empirical formulas [81], the per-unit-length values of the LCR components are found by

$$
\begin{aligned}
l &= 0.2 \ln\left(\frac{8H}{W} + \frac{W}{4H}\right) \mu\mathrm{H/m} = 2.90\,\mathrm{nH/cm} \\
c &= \varepsilon \left(\frac{W}{H} + \frac{\pi}{\ln\left[1 + \frac{2H}{T}\left(1 + \sqrt{1 + \frac{T}{H}}\right)\right]} - \frac{T}{4H} \right) = 1.08\,\mathrm{pF/cm} \\
r &= \frac{\rho_r}{WT} = 1.98\,\mathrm{m\Omega/cm}
\end{aligned}
\tag{4.7}
$$

As a result, without the circuit designer intentionally adding any LCR components to the power network, the PCB wiring itself resembles a typical "bias-tee" network — the DC voltage source V_{DD} provides the power to the functional circuit (M1), and the interference source $i_{if}(t)$ can "inject" disruptions into V_{DD}.

The interference source has the following definition:

$$
i_{if}(t) = A\ \exp\left[-\frac{\left(t - \left\lfloor \frac{t}{\tau_D} \right\rfloor \tau_D - \tau_{D0}\right)^2}{2\tau_L^2}\right] \cos\left[2\pi f_{ic}\left(t - \left\lfloor \frac{t}{\tau_D} \right\rfloor \tau_D - \tau_{D0}\right)\right] \tag{4.8a}
$$

$$
\cong A \sum_{n=0}^{N} \exp\left[-\frac{(t - n\tau_D - \tau_{D0})^2}{2\tau_L^2}\right] \cos\left[2\pi f_{ic}(t - n\tau_D - \tau_{D0})\right] \stackrel{\text{def}}{=} \widetilde{i_{if}}(t) \tag{4.8b}
$$

Compared to Equation 4.1, the additional sinusoidal term represents the carrier-frequency modulation at f_{ic}. The exact f_{ic} is determined to achieve maximum disruptions, which is done by performing a SPICE AC analysis of the whole circuit and finding the "resonance" frequency at $V_{DS,M1}$, which is typically 1–2 GHz. An example waveform is shown in Figure 4.10 for $\tau_D = 70\,\mathrm{ns}$, $\tau_{D0} = 20\,\mathrm{ns}$, $\tau_L = 1\,\mathrm{ns}$, and $f_{ic} = 1.8\,\mathrm{GHz}$.

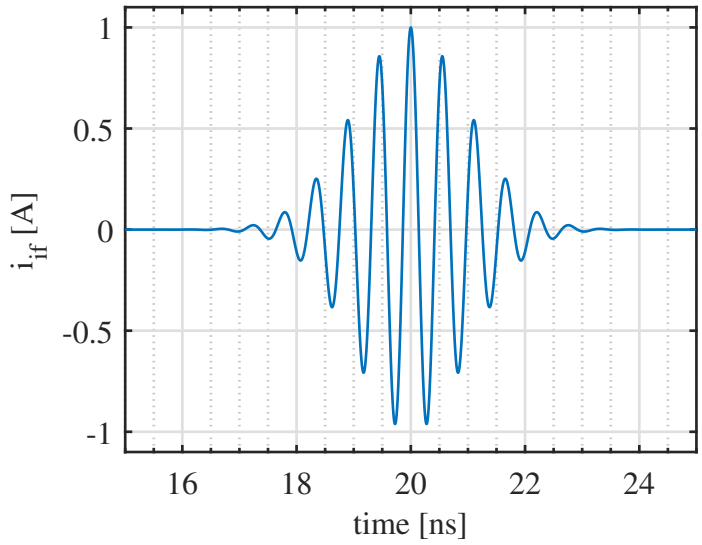


Figure 4.10: Time-domain waveform example for the Gaussian-enveloped sinusoidal pulses $i_{if}(t)$ used in Case #2 and #4. $A = 1\,\text{V}$, $\tau_D = 70\,\text{ns}$, $\tau_{D0} = 20\,\text{ns}$, $\tau_L = 1\,\text{ns}$ and $f_{ic} = 1.8\,\text{GHz}$. Only one "Gaussian packet" is shown in this plot.

When this type of interference is applied to the inverter circuit, the disrupted input/output waveforms are very similar to the ones in Case #1 shown in Figure 4.4. While the disruption "spikes" now become "symmetrical" or "balanced" around the expected voltage levels in V_{DD} of M1, we still anticipate possible transient bit errors that do not cause the circuit to latch up to an erroneous state indefinitely.

More interesting results are found when looking at the additional power consumption in M1 caused by the disruptions. Again, the average Joule power using terminal voltage and current waveforms are evaluated. The interference has a zero DC offset. Therefore, it should have no "long-term" influences by elevating the power rail voltage (V_{DD}). Each "packet" of the Gaussian-enveloped sinusoidal pulse has limited energy and power. The "injected" disruption level can be quantized by the average power of all pulses $\langle P_{inj} \rangle$ which can be calculated by using the three-sigma rule as below:

$$\langle P_{inj} \rangle = \frac{1}{6n\tau_L} \sum_{n=0}^{N} \int_{\tau_{D0}+n\tau_D-3\tau_L}^{\tau_{D0}+n\tau_D+3\tau_L} v_{if}(t)\ i_{if}(t)\, dt \tag{4.9}$$

where $v_{if}(t)$ and $i_{if}(t)$ are the instantaneous voltage and current measured at the "injection" port in the circuit in Figure 4.9.

Shown in Figure 4.11 is the additional power dissipation $\Delta\langle P_{M1}\rangle$ calculated using Equation 4.2 and 4.3, with and without the Soft Error model, drawn against the average injected power $\langle P_{inj}\rangle$ found with Equation 4.9. The time constants are $\tau_D = 70\,\text{ns}$, $\tau_{D0} = 20\,\text{ns}$, and $\tau_L = 1\,\text{ns}$. The interference's sinusoidal frequency is $f_{ic} = 1.8\,\text{GHz}$.

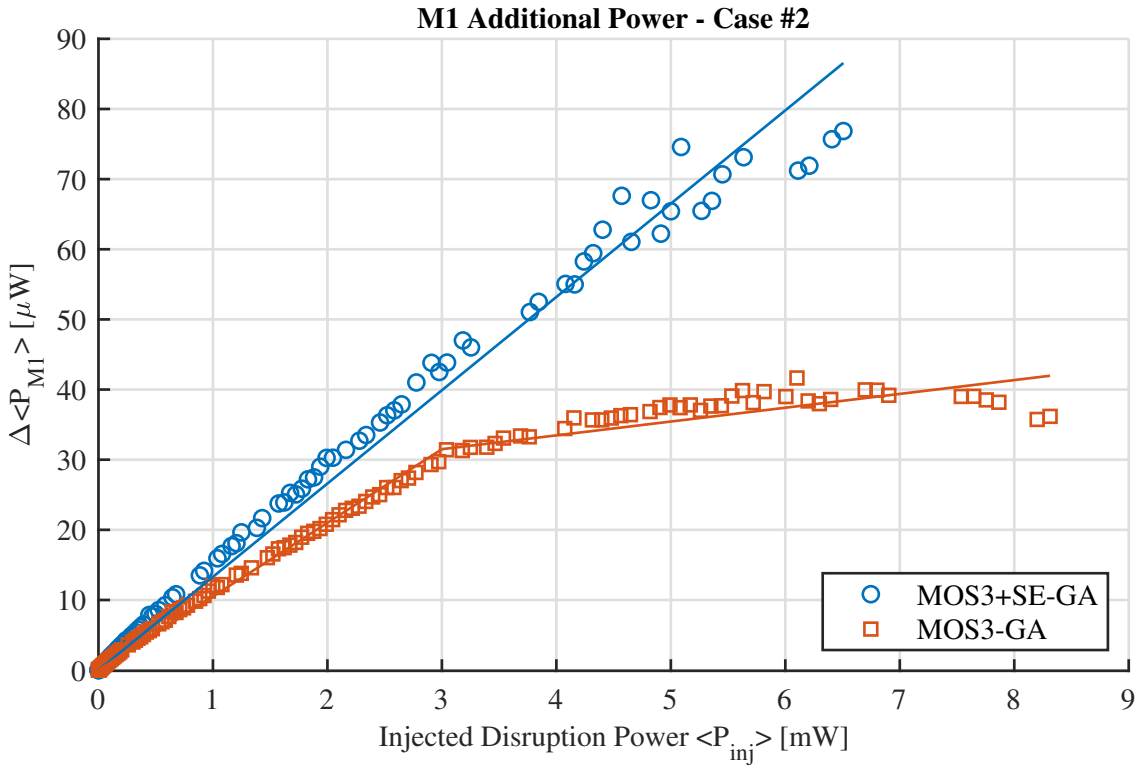


Figure 4.11: Additional power consumed by transistor M1 in Case #2. Results with and without the Soft Error model are labeled as "MOS3+SE-GA" and "MOS3-GA", respectively. The simulation data is shown by symbols, and the fit curves are shown by solid lines. The constant terms $\langle P_{M1,0}\rangle$ as in Equation 4.3 are deducted before drawing.

To summarize the "extra" vulnerability due to the Snapback phenomenon, we empirically fit the two sets of data with linear functions. The variable $\Delta\langle P_{M1}\rangle$ is in μW, and $\langle P_{inj}\rangle$ is in mW.

For the result group without Soft Error model, the dataset is divided into two segments with an arbitrary threshold at $\langle P_{inj}\rangle = 3\,\text{mW}$. The fitting results are:

with Soft Error model ("MOS3+SE-GA", fit line for the blue circles):

$$\Delta\langle P_{M1}\rangle = 13.3\langle P_{inj}\rangle + 16.2 \tag{4.10a}$$

without Soft Error model ("MOS3-GA", fit lines for the red squares):

$$\Delta\langle P_{M1}\rangle = \begin{cases} 10.5\langle P_{inj}\rangle & \langle P_{inj}\rangle \le 3\,\text{mW} \\ 1.97\left(\langle P_{inj}\rangle - 3\,\text{mW}\right) + 15.9 & \langle P_{inj}\rangle \ge 3\,\text{mW} \end{cases} \tag{4.10b}$$

The power consumption under no disruptions is $\langle P_{M1,0}\rangle = 16.2\,\text{mW}$ with the Soft Error model and 15.9 mW without it, same as in Case #1.

Overall, the additional power consumption increases as the disruption level increases. This is expected for the cases with and without our Soft Error model, because the volt-ampere product is generally higher.

With the Soft Error model enabled, the simulation always reports more additional power. It is interesting that for the non-Soft Error results, when $\langle P_{inj}\rangle$ becomes higher than the arbitrarily set threshold (3 mW), the increase in $\Delta\langle P_{M1}\rangle$ per increment in injection becomes much slower. The maximum difference is $\Delta\langle P_{M1}\rangle = 74.6\,\text{mW}$ with Soft Error when $\langle P_{inj}\rangle = 5.09\,\text{mW}$, which is higher than the non-Soft Error value 37.5 mW by 98.9 %. Thus, two separate fittings are performed for the non-Soft Error data, intercepting at $\langle P_{inj}\rangle = 3\,\text{mW}$. It corresponds to peak drain-source voltage $V_{DS} = 25.8\,\text{V}$ at $t = 2470.173\,\mu\text{s}$ and interference source amplitude $A = 79\,\text{mA}$, or $V_{DS} = 24.8\,\text{V}$ when the Soft Error model is enabled. On the other hand, when $\langle P_{inj}\rangle < 3\,\text{mW}$, about 30 % more extra power is reported from using the Soft Error model.

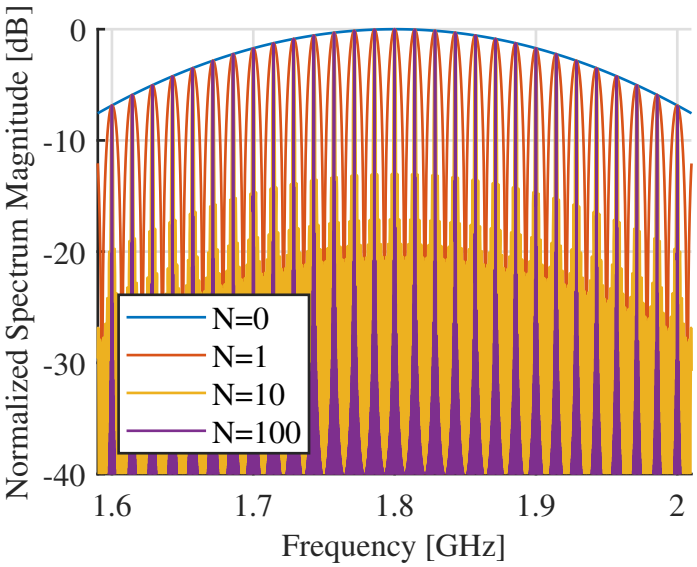


Figure 4.12: An example spectrum (normalized magnitude) for the disruption current source used in Case #2 and #4. The graph is generated by applying Fourier transform to the approximated form of the Gaussian-enveloped sinusoidal pulses $\left|\mathscr{F}\left\{\widetilde{i_{if}}(t)\right\}\right|(f)$ defined in in Equation 4.8b. In this case, $\tau_D = 70\,\text{ns}$, $\tau_{D0} = 20\,\text{ns}$, $\tau_L = 1\,\text{ns}$ and $f_{ic} = 1.8\,\text{GHz}$, which is used to generate the time-domain waveform shown in Figure 4.10. Various maximum N are used in the summation in Equation 4.11.

In a similar manner to Case #1, we look at the frequency components in the disruption source, the Gaussian-packet interference current source. Specifically, we use the approximated form $\widetilde{i_{if}}(t)$ given in Equation 4.8b. An example spectrum is shown in Figure 4.12. The Fourier transform of the approximated form is

$$\mathscr{F}\left\{\widetilde{i_{if}}(t)\right\}(f) = \frac{\sqrt{2\pi}A}{2f_L} e^{-j2\pi\tau_{D0}f} \sum_{n=0}^{N} \left\{ \exp\left[-2\pi^2\left(\frac{f-f_{ic}}{f_L}\right)^2 - j2n\pi\left(\frac{f}{f_D}\right)\right] + \exp\left[-2\pi^2\left(\frac{f+f_{ic}}{f_L}\right)^2 - j2n\pi\left(\frac{f}{f_D}\right)\right]\right\} \tag{4.11}$$

where $f_D = \tau_D^{-1}$ and $f_L = \tau_L^{-1}$. Its spectrum magnitude as shown in Figure 4.12 has a Gaussian peak at the carrier frequency f_{ic}, and zero DC offset which corresponds to the AC coupling effect. As the number of included "Gaussian packets" (N) increases, the spectrum converges to a collection of much narrower peaks separated by equal frequency intervals of f_D as can be measured from Figure 4.12, while the overall envelope still follows the Gaussian shape. This is due to the fact that the infinite sum of imaginary exponents

$\sum_{n=-\infty}^{+\infty} \exp\left[-j2\pi n\left({}^{f}/_{f_D}\right)\right]$, corresponding to a time-domain Delta function train, is the Fourier series expansion of a Delta function train in the frequency domain. The magnitude of all individual peaks follow the Gaussian envelope found in the case when $N = 0$; the center frequency is f_{ic}, and the cut-off frequencies are $f_{-3\,\mathrm{dB}} = f_{ic} \pm f_L\sqrt{\frac{\ln 2}{4\pi^2}}$.

For the same example case used before, when $f_{ic} = 1.8\,\mathrm{GHz}$ and $f_L = \tau_L^{-1} = (1\,\mathrm{ns})^{-1} = 1\,\mathrm{GHz}$, the bandwidth of the interference (full-width half-magnitude or FWHM) is found using the upper and lower cut-off frequencies as 0.26 GHz or 1.67–1.93 GHz. When the pulse-train separation is $\tau_D = 70\,\mathrm{ns}$, we have $f_D = \tau_D^{-1} = 14\,\mathrm{MHz}$, and the number of small Gaussian peaks inside this range is approximately $\mathrm{FWHM}/f_D = 0.26\,\mathrm{GHz} \div 14\,\mathrm{MHz} \approx 19$.

We can conclude that the interference is approximately a "single-frequency" (with a very narrow-band spectrum) stimulus to the circuit. When the functional circuit does not have an operational frequency band tuned to the interference frequency, the input and output signals are not likely to experience resonance or stimulated oscillation. However, as can be seen from the simulation results, power rail disruptions of this type could still indirectly affect the output, such as by affecting the bias point after causing resonance in the PCB traces.

4.3 Case Study #3: Narrow-Band RF Amplifier under Unipolar Interferenece on Power Line

An analog amplifier is designed with CoolSpice Package using only the N-MOSFET used in previous cases, and passive RLC components. The circuit diagram is shown in Figure 4.13, including the source and a passive load, but omitting the disrupted power supply (V_{DD}), which would be the same as in Case #1. The amplifier consists of two cascaded stages: a common-source amplifier succeeded by a source follower. The internal loads at the drain of M1 and source of M2 are tuned RLC band-pass filters. The passive RLC networks at M1 and M2's gates are tuned band-pass filters which also serve as biasing networks (the resistors may be replaced by self-biased transistors). When the matched input source with a $50\,\Omega$ internal impedance has a peak-to-peak amplitude of $v_{in}(t) = 12.6\,\text{mV}$, and the output port $v_{out}(t)$ is matched to a $1\,\text{k}\Omega$ load, this two-stage amplifier provides a small-signal gain of 12.9 dB at central frequency $f_{ac} = 7.02\,\text{MHz}$ and a bandwidth of $BW = \pm 0.01\,\text{MHz}$ (half-width half-magnitude).

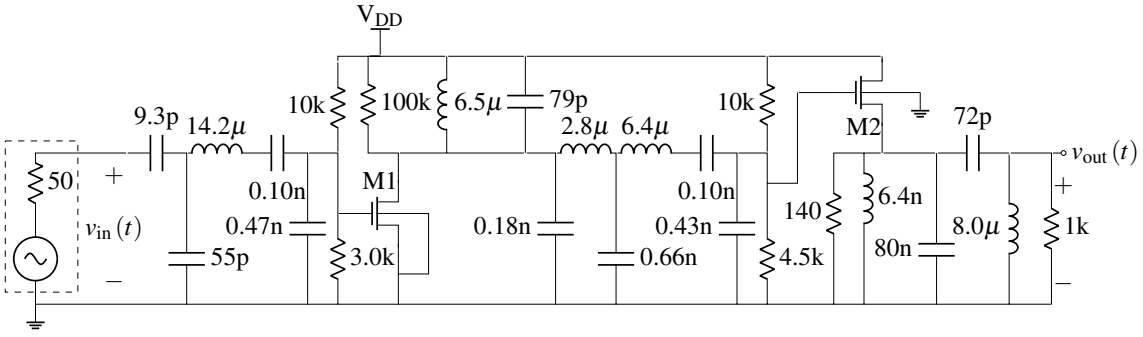


Figure 4.13: Circuit diagram of a two-stage narrow-band RF amplifier used in Case #3 and #4. The disruptions in the voltage rail (V_{DD}) are omitted in the figure.

In Figure 4.14, we examine the waveforms of the power, input, output, and MOSFET terminals when power supply disruptions are present. The single-peaked Gaussian pulses are not present until simulation time $t = 40\,\mu\text{s}$, so that the amplifier has reached the steady state before the disruptions start. The Gaussian peaks in V_{DD} in this example have amplitude $A = 7.4\,\text{V}$. The time constants are $\tau_D = 70\,\text{ns}$, $\tau_{D0} = 20\,\text{ns}$, and $\tau_L = 3\,\text{ns}$.

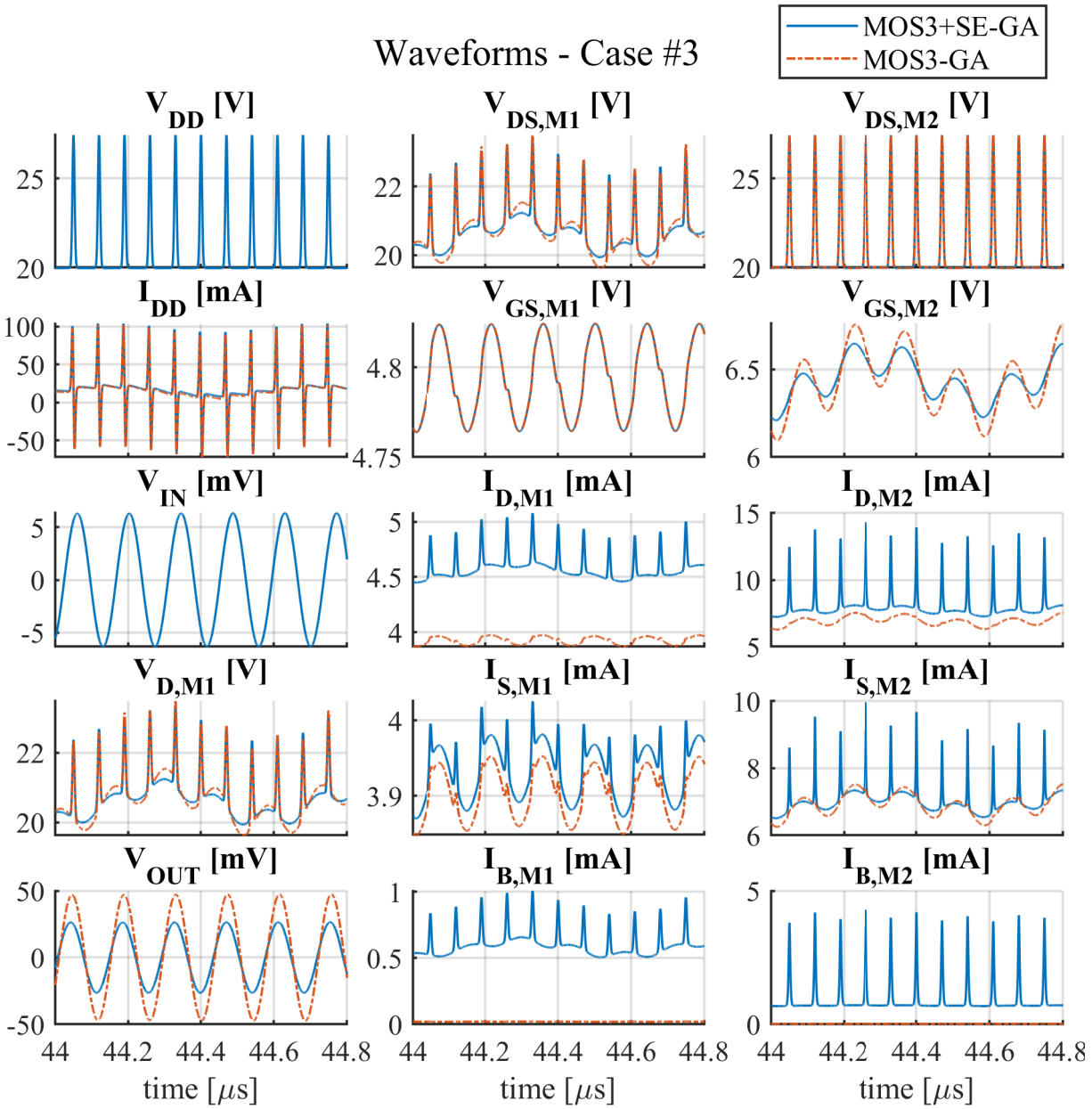


Figure 4.14: Waveforms of Case #3. The Soft Error model results ("MOS3+SE-GA") and basic MOSFET model results ("MOS3-GA") are compared. The disruption amplitude is $A = 7.4\,\mathrm{V}$. $\tau_D = 70\,\mathrm{ns}$, $\tau_{D0} = 20\,\mathrm{ns}$, and $\tau_L = 3\,\mathrm{ns}$.

The amplifier's input signal ($V_{GS,M1}$) is barely interrupted since the out-of-band interference (14 MHz base frequency and higher-order harmonics; see the Fourier analysis in Case #1) is mostly filtered out by the shunt capacitors and serial inductor in the input BPF. However, the drain voltage $V_{DS,M1}$ is much more affected since the BPF load is effectively shorted out-of-band. As a consequence, terminal current increases are observed.

Also, the input signal of M2 ($V_{GS,M2}$) contains a beat frequency of approximately 1.7 MHz, which is absent when no power supply disruptions are present. By comparing to Case #4, it is believed that since the unipolar single spikes have a wide-band spectrum and a DC offset, the interference may have caused a resonance at this frequency, much lower than the operation frequency of the amplifier. However, because of the load BPF connected to M2's source, the below-band beat frequency is removed by the shunt inductor from the output signal.

We also observe a slight change in the peak-to-peak amplitude of the amplified signals. By comparing the first stage's input and the second stage's output voltages ($V_{DS,M1}$ and $V_{S,M2}$) with and without the Soft Error model, a decrease in the amplitude gain is found. This will be analyzed further shortly.

By comparing the time each waveform reaches its non-disruption peaks, it is found that by applying the Soft Error model, little to no additional time delay or phase lag is introduced. In $V_{DS,M1}$, the RMS value of time delay is 4.12 ns among the seven "signal peaks" within the simulation time $t = 44$–$45\,\mu s$, while a signal cycle is $T = (7\,\text{MHz})^{-1} = 143\,\text{ns}$. The $V_{GS,M2}$ and $V_{S,M2}$ waveforms have 4.64 ns and 3.70 ns for the same quantity, respectively. Therefore, about 3 % of error in phase delay in the observed voltages could be related to the application of the Soft Error model in this circuit.

Next, we summarize the two changes in the amplifier's behavior versus various levels of disruption. In Figure 4.15, the average additional power dissipation of M1 and M2, individually, are evaluated using Equation 4.3. For M1:

$$\langle P_{M1} \rangle = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \left[i_{D,M1}(t)\ v_{D,M1}(t) - i_{S,M1}(t)\ v_{S,M1}(t) - i_{B,M1}(t)\ v_{B,M1}(t) \right] dt \quad (4.12)$$

in simulation time $t = 44$–$45\,\mu s$. The same applies for M2.

The Soft Error model reports more drawn power for both transistors, as can be predicted by the significantly higher I_D and I_B shown in Figure 4.14. The empirical exponential or polynomial fittings are drawn in solid lines.

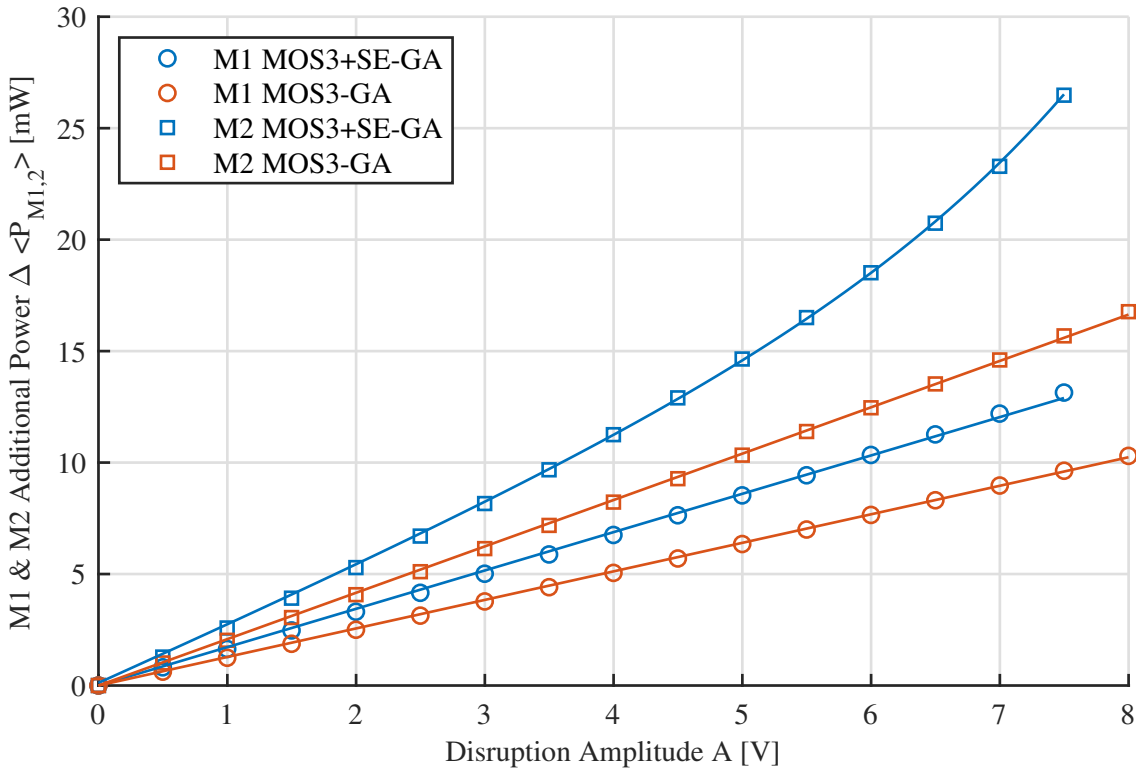


Figure 4.15: Additional power consumed by transistors M1 and M2 in Case #3. Results with and without the Soft Error model are labeled as "MOS3+SE-GA" and "MOS3-GA", respectively. The amplitude of the unipolar Gaussian pulses A (x-axis) in the simulation data has a uniform interval of 0.1 V. The simulation data points (symbols) and fit curves (solid lines) represent the *additional* power $\Delta\langle P_{M1} \rangle$ and $\Delta\langle P_{M2} \rangle$ in Equation 4.3.

The empirically fit expressions are listed with A in V and $\Delta\langle P_{M1}\rangle$, $\Delta\langle P_{M2}\rangle$ in mW as below:

First stage (M1):

with Soft Error model ("MOS3+SE-GA", fit line for the blue circles):

$$\Delta\langle P_{M1}\rangle = 1.72\,A \tag{4.13a}$$

without Soft Error model ("MOS3-GA", fit line for the red circles):

$$\Delta\langle P_{M1}\rangle = 1.28\,A \tag{4.13b}$$

Second stage (M2):

with Soft Error model ("MOS3+SE-GA", fit line for the blue squares):

$$\Delta\langle P_{M2}\rangle = 0.120\,\exp(0.551A) + 2.54\,A \tag{4.13c}$$

$$\text{(alternatively) } \Delta\langle P_{M2}\rangle = 0.160\,A^2 + 2.19\,A \tag{4.13d}$$

without Soft Error model ("MOS3-GA", fit line for the red squares):

$$\Delta\langle P_{M2}\rangle = 2.08\,A \tag{4.13e}$$

The constant terms $\langle P_{M1,0}\rangle$ and $\langle P_{M2,0}\rangle$ as in Equation 4.3 represent the no-disruption power consumption. Their values with the Soft Error model are $\langle P_{M1,0}\rangle = 82.7\,\text{mW}$ and $\langle P_{M2,0}\rangle = 144\,\text{mW}$. Without the Soft Error model (when only the baseline model is used), they are $\langle P_{M1,0}\rangle = 72.3\,\text{mW}$ and $\langle P_{M2,0}\rangle = 130\,\text{mW}$.

Overall, the additional power consumption increases as the disruption level increases. This is expected for the cases with and without our Soft Error model, because the volt-ampere product is generally higher.

It is worth noting that: (1) in the linear fittings, the increase in the extra power $\Delta\langle P\rangle$ per increment in the disruption level A is about 30 % higher when the Soft Error model is included; (2) the source follower (M2) has an either quadratic or exponential trend in

$\Delta\langle P_{M2}\rangle$ with the Soft Error model, different than M1 and in Case #1.

Besides the additional power consumption due to the interference, the function of this circuit is also affected. In Figure 4.16, the change in overall small-signal gain given by

$$\Delta G = G - G_0 \tag{4.14}$$

is shown versus levels of disruptions.

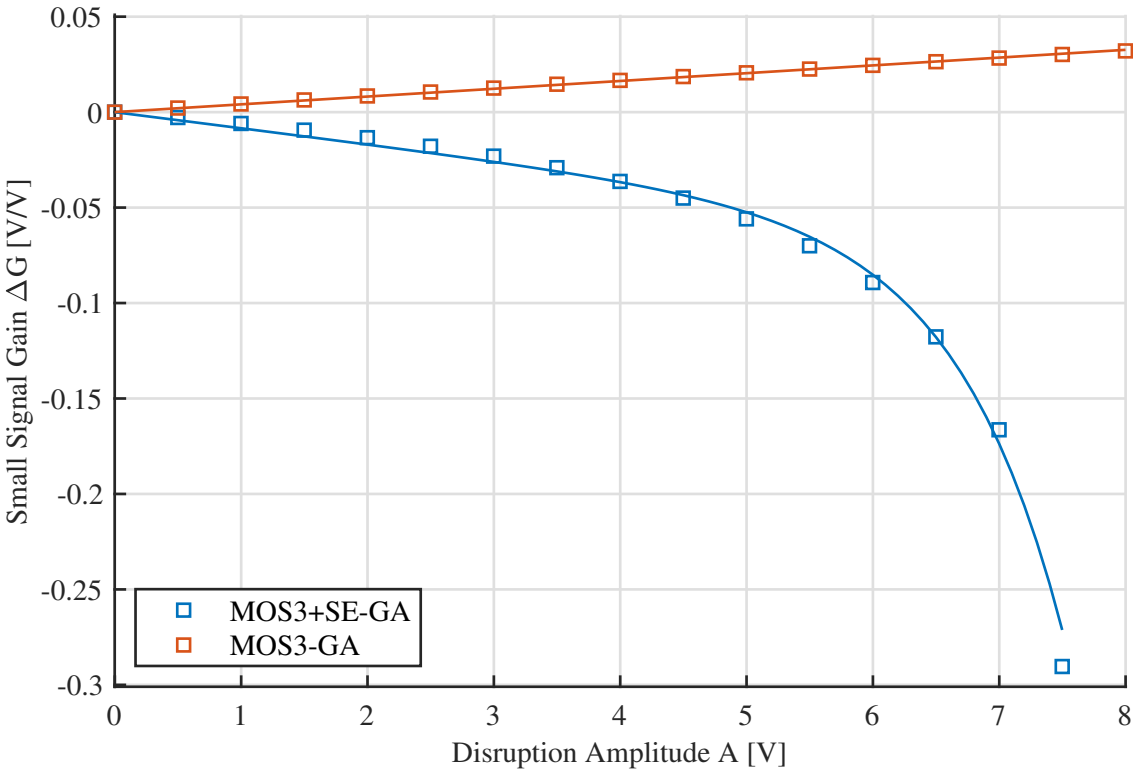


Figure 4.16: Change in the small signal gain ΔG in Case #3. Results with and without the Soft Error model are labeled as "MOS3+SE-GA" and "MOS3-GA", respectively. The amplitude of the unipolar Gaussian pulses A (x-axis) in the simulation data has a uniform interval of 0.1 V. The simulation data points (symbols) and fit curves (solid lines) represent the *change* in the gain.

The gain G is defined as the ratio between the input and output signal peak-to-peak amplitudes. G_0 is the gain with no disruptions ($A = 0$). The empirically fit expressions are listed below with A in V and ΔG in V/V:

with Soft Error model ("MOS3+SE-GA", fit line for the blue squares):

$$\Delta G = -2.78 \times 10^{-5} \exp(1.19A) - 8.36 \times 10^{-3} A \tag{4.15a}$$

without Soft Error model ("MOS3-GA", fit line for the red squares):

$$\Delta G = 4.08 \times 10^{-3} A \tag{4.15b}$$

The constant term G_0 representing the no-disruption gain is 4.43 V/V with the Soft Error model and 7.44 V/V without it. While it is helpful for a circuit designer to implement and calibrate the additional capacitors from the activated source-body junction and the impact ionization-induced charges, they are not implemented in our Soft Error model at present since our calibration work is mainly on the large-signal behaviors. Therefore, the difference in G_0 is most likely due to the slight impact ionization drain-body current.

Despite the apparent lower gain with no disruptions when the Soft Error model is included (40.5 % or −4.50 dB lower), it is surprising that the gain decreases as the level of disruptions increases, while the basic, non-Soft Error model reports a slightly increasing trend in contrast.

4.4 Case Study #4: Narrow-Band RF Amplifier under Symmetric Interference on Power Line

The same amplifier as in Case #3 is tested with the interference introduced between the true DC power supply and the "power rail" V_{DD} of the functional circuit. The circuit components related to the interference are the same as those in Case #2 (r_1, l_1, c_1, r_2, l_2, c_2, L_{if}, and $i_{if}(t)$ in Figure 4.9), except the power rail wire lengths before and after the interference injection point (i_{if}) are now 6 cm and 0.1 cm, respectively, and the interference's sinusoidal frequency is $f_{ic} = 1.25\,\text{GHz}$.

The waveforms when the disruption level of $i_{if}(t)$ in Equation 4.8 is $A = 2.2\,\text{mA}$ are shown in Figure 4.17. Similar to Case #3, increased I_D, I_S and I_B and spikes in terminal voltages and currents are found, but the second stage's input signal ($V_{GS,M2}$) does not contain beat frequencies. It is believed that because the Gaussian-enveloped sinusoidal interference has a narrow spectrum with no DC offset, the potential resonance frequency found in Case #3 has not been excited.

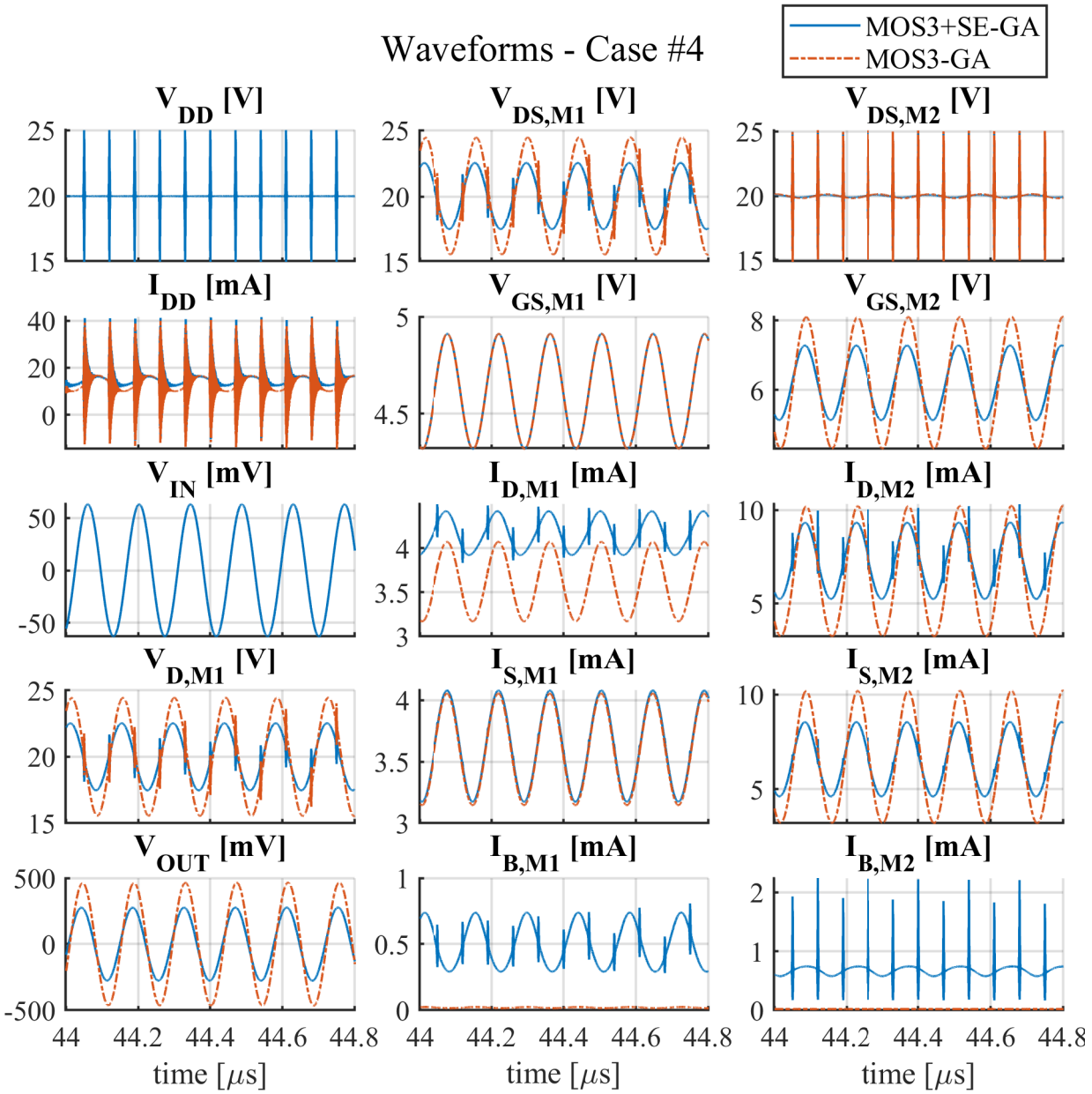


Figure 4.17: Waveforms of Case #4. The Soft Error model results ("MOS3+SE-GA") and basic MOSFET model results ("MOS3-GA") are compared. The disruption amplitude is $A = 2.2\,\text{mA}$. $\tau_D = 70\,\text{ns}$, $\tau_{D0} = 20\,\text{ns}$, $\tau_L = 1\,\text{ns}$ and $f_{ic} = 1.25\,\text{GHz}$.

In Figure 4.18, the additional power consumption caused by disruptions in M1 and M2 ($\Delta\langle P_{M1}\rangle$, $\Delta\langle P_{M2}\rangle$) is evaluated in simulation time $t = 44$–$45\,\mu s$ by Equation 4.12 and 4.3, and drawn against the average interference injection power $\langle P_{inj}\rangle$ found by Equation 4.9.

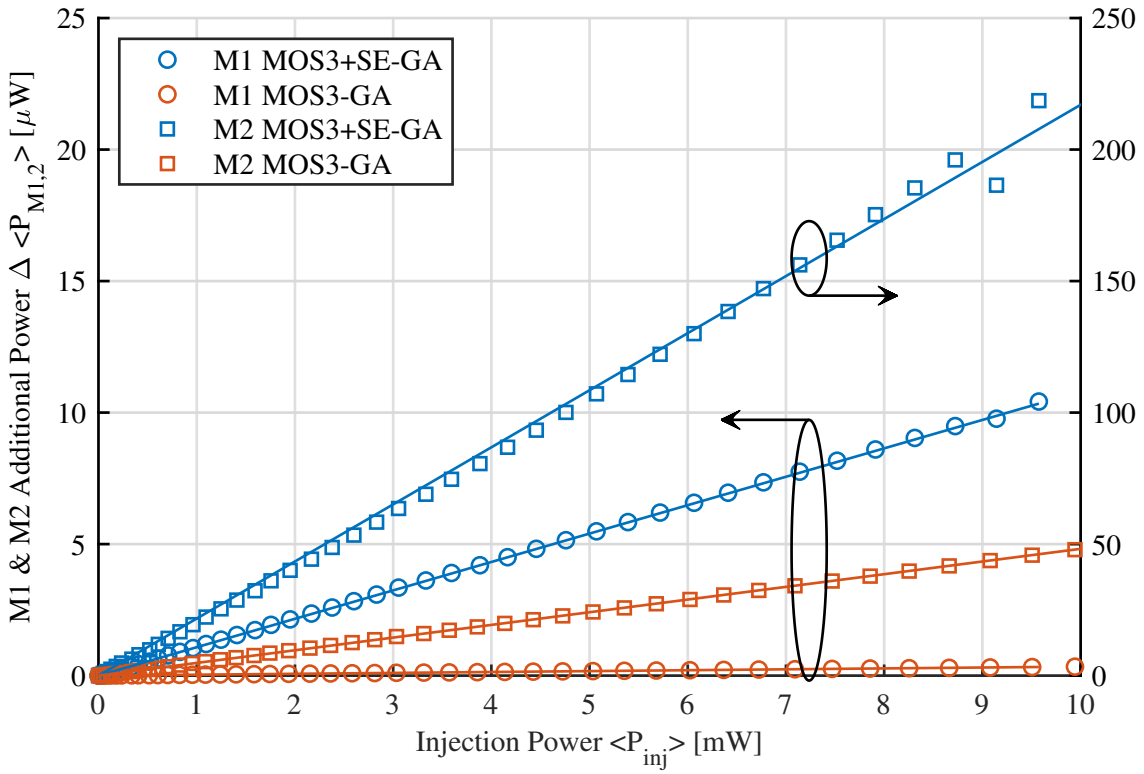


Figure 4.18: Additional power consumed by transistors M1 and M2 in Case #4. Results with and without the Soft Error model are labeled as "MOS3+SE-GA" and "MOS3-GA", respectively. The simulation data points (symbols) and fit curves (solid lines) represent the *additional* power $\Delta\langle P_{M1}\rangle$ and $\Delta\langle P_{M2}\rangle$ in Equation 4.3 versus the average injected power $\langle P_{inj}\rangle$.

The empirically fit expressions are listed below with $\langle P_{inj} \rangle$ in mW and $\Delta\langle P_{M1} \rangle$, $\Delta\langle P_{M2} \rangle$ in μW.

First stage (M1):

with Soft Error model ("MOS3+SE-GA", fit line for the blue circles):

$$\Delta\langle P_{M1} \rangle = 1.08 \langle P_{inj} \rangle \tag{4.16a}$$

without Soft Error model ("MOS3-GA", fit line for the red circles):

$$\Delta\langle P_{M1} \rangle = 0.0342 \langle P_{inj} \rangle \tag{4.16b}$$

Second stage (M2):

with Soft Error model ("MOS3+SE-GA", fit line for the blue squares):

$$\Delta\langle P_{M2} \rangle = 21.7 \langle P_{inj} \rangle \tag{4.16c}$$

without Soft Error model ("MOS3-GA", fit line for the red squares):

$$\Delta\langle P_{M2} \rangle = 0.482 \langle P_{inj} \rangle \tag{4.16d}$$

The constant terms $\langle P_{M1,0} \rangle$ and $\langle P_{M2,0} \rangle$ as in Equation 4.3 represent the no-disruption power consumption. Their values with the Soft Error model are $\langle P_{M1,0} \rangle = 82.9\,\text{mW}$ and $\langle P_{M2,0} \rangle = 145\,\text{mW}$. Without the Soft Error model (when only the baseline model is used), they are $\langle P_{M1,0} \rangle = 71.4\,\text{mW}$ and $\langle P_{M2,0} \rangle = 132\,\text{mW}$.

Overall, the additional power consumption increases as the disruption level increases. This is expected for the cases with and without our Soft Error model, because the volt-ampere product is generally higher.

The linear fittings show the increasing trend in $\Delta\langle P_{M1} \rangle$ and $\Delta\langle P_{M2} \rangle$ when $\langle P_{inj} \rangle$ increases. The linear coefficient in M1's fitting, representing the increment in additional power per increment in injection power, is 31.6 times as high after the Soft Error model is enabled, while in the case of M2, it is 45.0 times as high.

In Figure 4.19, the changes in gain ΔG found by Equation 4.14 are plotted against the injection power $\langle P_{inj} \rangle$.

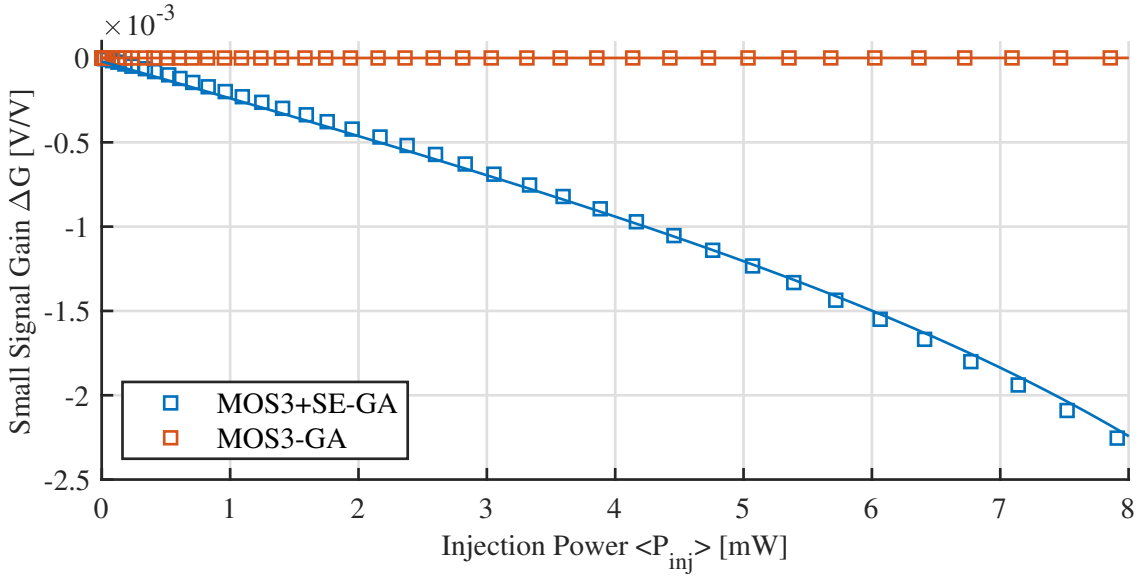


Figure 4.19: Change in the small signal gain ΔG in Case #4. Results with and without the Soft Error model are labeled as "MOS3+SE-GA" and "MOS3-GA", respectively. The simulation data points (symbols) and fit curves (solid lines) represent the *change* in gain ΔG in Equation 4.3 versus the average injected power $\langle P_{inj} \rangle$.

The empirically fit expressions are listed below with $\langle P_{inj} \rangle$ in mW and ΔG in V/V.

with Soft Error model ("MOS3+SE-GA", fit line for the blue squares):

$$\Delta G = -2.04 \times 10^{-5} \exp\left(0.418 \langle P_{inj} \rangle\right) - 2.08 \times 10^{-4} \langle P_{inj} \rangle \tag{4.17a}$$

without Soft Error model ("MOS3-GA", fit line for the red squares):

$$\Delta G = 8.04 \times 10^{-8} \langle P_{inj} \rangle \tag{4.17b}$$

The constant term G_0 representing the no-disruption gain is 4.40 V/V with the Soft Error model and 7.40 V/V without it. While it is helpful for a circuit designer to implement and calibrate the additional capacitors from the activated source-body junction and the impact ionization-induced charges, they are not implemented in our Soft Error model at present since our calibration work is mainly on the large-signal behaviors. Therefore, the difference in G_0 is most likely due to the slight impact ionization drain-body current.

Overall, the change in amplitude gain ΔG is small compared to the base-line value G_0. Nevertheless, it is worth noticing the differences once the Soft Error model is enabled. The basic MOSFET model reports a minimal increasing trend in gain as $\langle P_{inj} \rangle$ increases. In contrast, when the Soft Error model is included, a non-linear decreasing trend ΔG is reported. The empirical linear and exponential fittings are listed in Figure 4.19, where ΔG is in V/V (linear scale), and $\langle P_{inj} \rangle$ is in mW.

Chapter 5: EMI-Induced Hard Failure Vulnerabilities: Oxide Breakdown Mechanisms and Rapid, Permanent Breakdown Experiments

Hard Failures refer to permanent damages to MOSFET devices. They lead to device characteristic changes, which may be catastrophic to circuit functions. One of the prominent types of Hard Failures is gate dielectric breakdown (BD). In Section 5.1, the mechanisms leading to permanent dielectric BD are reviewed. The dielectric material studied in this work is SiO_2 ("oxide"), but the concepts can be adapted to other dielectric structures and materials. Depending on the level of electrical stress, which can be caused by transient disruptions under EMI conditions, the permanent BD process may happen very fast, i.e., almost instantaneously so long the disruption occurs (Rapid BD or RBD); in other cases, the BD process may be slower and affect the device's and the circuit's projected lifetime (Long-Term BD or LBD).

We carried out rapid oxide breakdown experiments using our own MOSFET devices. We fabricated standalone N-MOSFET devices at the NanoCenter FabLab, University of Maryland, described in Section 5.2. We performed gate stress tests on these test devices. In Sections 5.3 and 5.4, RBD events under DC and low-speed transient pulse stress conditions are observed, and the data is collected and analyzed from statistical perspectives. The knowledge of RBD conditions is later used to build an empirical RBD circuit model described in Section 6.4.

5.1 Review of Oxide Breakdown Mechanisms

Gate oxide breakdown (BD) as a reliability factor has been a topic of interest for decades since the destruction of individual transistors severely affects the lifetime of integrated circuits. While the exact physical mechanism leading to the breakdown is still under

debate, several physics-based models are well supported by experimental data [14, 45, 82–84].

Although extrinsic breakdown mechanisms such as impurity defects and layout related vulnerabilities can be alleviated by improving the already complicated fabrication process, intrinsic mechanisms related to the basic physical structure of MOSFETs are of particular interest to us.

The BD mechanism described by Lombardo et. al. [14] consists of three stages, including: (i) the gradual, accumulative wear-out stage, due to voltage stress applied on the gate; (ii) the on-set event of breakdown, usually after the formulation of a continuous path capable of massive current conduction; and (iii) the post-breakdown damage, leading to the non-recoverable device corruption. The reported experimental results suggest that the physical mechanisms behind the "fast" or "hard" BD and the "slow" or "soft" BD are similar, while the major difference is the level of stress determining the time before the initial BD events and the level of post-BD damages. After reviewing the three stages in the oxide BD process, we introduce the terms "Rapid BD" (RBD) and "Long-Term BD" (LBD) used in this work.

1) Wear-Out Stage

In the wear-out stage, defects are created by tunneling current, which may be of different types depending on the oxide thickness and the applied voltage between gate and substrate [83]. When a high voltage is applied across a thick oxide, the conduction band of the oxide drops below that of the gate poly-Si (n-type) or substrate Si (p-type), which is 3.1 eV below the former[16] under equilibrium at the interface. Under this condition, *Fowler-Nordheim tunneling* happens, where the conduction-band electrons from the Si cathode tunnel into the oxide's conduction band, and proceed towards the anode. When the field is not high enough for the above mechanism, but the oxide is thin (usually below 2 nm), *direct quantum tunneling* is also possible. The third type of leakage comes from *thermal*

[16]The conduction band potentials of silicon dioxide and silicon are 0.95 eV and 4.05 eV below vacuum level, respectively. [85]

emission, where electrons overcome the 3.1 eV barrier between the conduction bands at the Si-SiO_2 interface.

When gate leakage current is present, no matter which mechanism it originates from, the electrons as mobile carriers are capable of causing structural damage, by assisting with the creation of defect sites in the oxide. The defects are neutral traps where the band gap is smaller than normal so that electrons can tunnel into these traps before continuing on across the rest of the oxide region, forming trap-assisted tunneling current. The mechanism that leads to the creation of such defects is still under investigation and described by different models [14, 45, 82–84].

The *hydrogen release model* [82] and *anode hole injection model* [83] believe that hot carriers (mainly electrons) having already tunneled through the oxide can induce the release of ionized hydrogen atoms or mobile holes at the silicon anode. Generated particles diffuse [86] (or possibly drift, depending on the transient field by applied bias) towards the cathode, through the oxide. During the transport process, defect sites are created. In the hydrogen release model, it is believed that a hydrogen bridge is formed [14], which consists of a hydrogen atom, released at the Si-SiO_2 interface, and bonds with two neighboring silicon atoms in place of an oxygen vacancy. These bridges increase the possibility of low-energy electrons tunneling through, hence the process is called *trap-assisted tunneling*.

It is also believed that impact ionization-induced hot holes are capable of causing the hole-induced trap generation (anode hole injection model). The generated holes, initially in the channel (for p-type) or gate metal (for n-type) depending on the polarity of MOSFET, tunnel back into the oxide only with a small probability [84]. These holes can be "captured" in Si-O bonds, reducing the ionic bond energy, and lowering the activation energy required to break the bond [83]. The captured hole itself is not involved in the transition from sp^3 orbital to sp^2 (broken-bond state), nor does it contribute energy (causing atomic displacement) by colliding with the lattice due to the dramatic difference between electron and atom mass. Instead, it catalyzes the reaction, although it is possible that the

captured hole will recombine with an electron and thus be consumed.

Apart from the two models above which relate the generation of defect sites to extraneous particles, the *E-model* or *thermo-chemical model* [45] states that the degradation rate k is determined by the Arrhenius equation,

$$k \propto \exp\left(-\frac{\Delta H - a\,E}{kT}\right) \tag{5.1}$$

which is written in terms of the activation energy ΔH related to the chemical reaction, the electric field E, and temperature T in the oxide. It is assumed in this model that "defective sites" randomly exist in a thermally grown amorphous SiO_2 network. The defective sites include abnormal atomic configurations that are much different than the crystalline SiO_2 structure (α-quartz), such as extreme bond angles between Si and O atoms and missing atoms causing Si-Si or O-O bonds. The concentration of these defects in a pristine oxide (that has never experienced electrical or thermal stress) is typically spatially uniform and can be measured experimentally [87]. The applied field E causes polarization, and both the applied and polarization fields together (as the term aE where a includes the dipole field due to polarization) add stress to the localized defective bonds. Meanwhile, the molecule with the distorted bond receives thermal energy (kT) from the lattice. The thermochemical model also states that the reaction rate depends on the activation energy ΔH to break the defective bond, which is effectively reduced by the external and polarization field (aE, combined). Interaction with holes or hydrogen can be addressed by an additional reduction in the activation energy. The reaction can also be accelerated by increasing the lattice temperature. Although the validity of the E-model as it relates to the real physical mechanism is under debate, it provides a good fit to certain experimental data, and it is a simple way to estimate the rate of the degradation and therefore the time to failure of the device.

2) On-set of Oxide BD Event

When the defects accumulate to a critical density, a percolation path through the oxide is formed, consisting of layers of localized defect sites. Each site has an effective cross

section of around 3 nm, much larger than its physical dimensions [14]. The created traps are now able to assist carrier transport through the entire oxide region. Large conduction current is observed, starting the breakdown event.

Before moving on to the post-BD effects, we would like to make a few comments:

- Ultra-thin SiO_2

For extremely thin oxides of 5 nm and thinner, a percolation path is not needed since a single defect site is enough to cause significant trap-assisted current. Also, direct tunneling is more problematic for extremely thin oxide layers, since the de-Broglie wavelength of the channel electron wavefunctions in the SiO_2 band gap is around 0.1 nm (see Section 6.3 and Equation 6.35).

Additionally, ballistic transport of electrons becomes highly possible in this range of oxide thickness [88]. Evidence was found that in thin oxides, the defect generation is solely dependent on the applied voltage and independent of the oxide thickness. The gate voltage threshold for defect generation is around 5 V, stemming from the 2 eV trap creation (as can be seen from the hole capture process) and the 3 eV potential barrier for FN tunneling. This threshold voltage is not a "hard" limit, meaning that exponential dependency can be observed when the applied voltage is below threshold. The reason could be that the the electron population versus energy distribution has an exponential-like high-energy tail.

In conclusion, the intrinsic oxide breakdown mechanism still exists for ultra-thin oxides, even under very low bias conditions.

- High-K Materials

High-K materials are of interest to small-scale devices thanks to their higher relative dielectric constants and hence reduced effective oxide thickness, which means similar electrical performance to ultra-thin SiO_2 gates can be achieved with a physically thicker dielectric. Ultra-thin SiO_2 films of less than 2 nm suffer from increased tunneling leakage

current [89]. By using high-K materials, this issue is mostly alleviated since a very low effective oxide thickness can be achieved with a relatively thick physical dielectric.

This is an advantage of using high-K materials, although the situation is often complicated by the fact that high-K dielectrics generally tend to have smaller conduction band offsets at the interface, so more thermionic emission is possible.[17]

Investigations on gate stacks consisting of thin HfO_2 and interfacial SiO_2 films have been reported [34, 91–93]. Despite the thicker high-K layers, experimental and simulation results show that gate leakage still exists in these gate structures. The progressive breakdown behavior is very similar to the SiO_2-only structure. The formation of localized percolation path is the key event leading to excessive gate current and non-reversible breakdown [93]. Despite the defect formation energy of 5 eV or even higher in bulk HfO_2, a group reported that for a thin HfO_2 film it was possible to have trap creations under the influence of the interfacing SiO_2, which can dramatically reduce this energy [92]. In fact, another group found that the rate of progressive trap creation (and breakdown) was controlled by the SiO_2 degradation [91].

From the electrical field's perspective, the BD condition for HfO_2 is similar to SiO_2. Experimental data suggests that HfO_2 dielectrics may BD instantaneously at around 8–10 MV/cm [47] or in minutes to hours at a moderate field of about 6 MV/cm [48].

3) Post-Breakdown Damage

After the breakdown event, the gate leakage current grows rapidly. The leakage current may come from the constant field applied across the gate oxide or the discharge of the MOS capacitor. The former case is more likely to lead to progressive growth of defects starting from the initial conducting site (either a percolation path or, for ultra-thin oxide, a single-site defect), gradually increasing the leakage current; the latter case could lead to the structural breakdown in a short period of time due to generated heat. The uncontrolled

[17] For example, at a direct intefrace between Si and HfO_2, the conduction band bottoms are 4.05 eV [85] and 1.88 eV below vacuum level [90], respectively, and therefore the conduction band offset is 2.17 eV [90], which is less than 3.1 eV for a Si-SiO_2 interface.

gate current or the non-recoverable characteristics change marks the final breakdown of the device.

4) Rapid BD vs Long-Term BD

While SiO_2 in the gate oxide can break down under stress conditions as any other insulating material does, it is worth noting the difference between the "rapid" breakdown (RBD), which happens very fast under extreme conditions such as a very high field (generally $\sim 10\,\mathrm{MV/cm}$), and a "long-term" (or "progressive") breakdown (LBD), which can happen under less severe conditions but in a much longer time. The terms "rapid" and "long-term" generally refer to the time scale and severeness of damage of these events in the oxide structure. In the literature they are often referred to as "hard" and "soft" breakdown, respectively. An alternative way is used to address the same issues to avoid confusion with other content in this work.

The difference between "rapid" BD (RBD) and "long-term' BD (LBD) is often confusing and varies depending on the choice of criteria, such as the measured time to BD, post-BD resistance, or the rate of gate leakage current growth. According to an experiment reported by Lombardo et. al. [14], the two BD mechanisms are related at the physical level, and the major difference is the time scale of BD events compared to the operational speed of test equipment, and the post-BD damage.

In the experiment, gate capacitors of size $1 \times 10^4\,\mu\mathrm{m}^2$ and various thickness were stressed under excessively high field ($> 10\,\mathrm{MV/cm}$), which was guaranteed to cause instant RBD in all oxides, in the time scale of 10 ns to 100 ns. High terminal voltage was held constant prior to the BD event by using a parallel capacitor. After the initial BD event, the terminal current rapidly increased and was limited at the compliance level, until the stress was removed and the test ended (however, the parallel capacitor can discharge without compliance). In the post-BD inspection, the gate polysilicon was found damaged with local melting in a particular pattern, which appeared as a self-avoiding random walk, indicating lateral propagation of the melted sites, which in turn are related to BD events.

The propagation velocity was calculated, and it was faster than the heat diffusion velocity, showing it was not possible for the heat produced by the previous BD site to cause BD and melting of the next BD site. Also, by solving the time-dependent heat equation of the local area, the thermionic emission current under transient temperature increase was not enough to cause local melting in the observed time span. Therefore, the melting could only be a result of the RBD event, rather than the cause. The propagation velocity was also faster than the speed of sound, suggesting that any possible structural damage due to atomic dislocation could not be the cause of such propagation. Instead, the propagation velocity was close to that of electron diffusion in the gate polysilicon. This can be better explained with the proposed model by Lombardo et.al., which is summarized below and is similar to the mechanism described previously.

5) Summary

The model can be summarized as follows. Before the eventual BD event, there is a build-up of defect spots. Because only a fraction of the defective sites can conduct, the initial BD event occurs after the accumulation of defective sites has reached a substantial level. (For ultra-thin oxides, evidence was found that the random walk ceased, indicating a lack of defect build-up next to the initial BD site.) At this time, the surrounding defective sites ("weak spots") that are almost complete forming conductive paths can become conductive under the influence of the initial site, already conducting. That requires the distance between such localized weak spots, in proportion to the inverse of the defect density, to be no larger than that could be reached by the diffusion of the electrons in the initial conductive site. Then, a chain reaction can happen, leading to a self-avoiding random walk, or the spread of BD events.

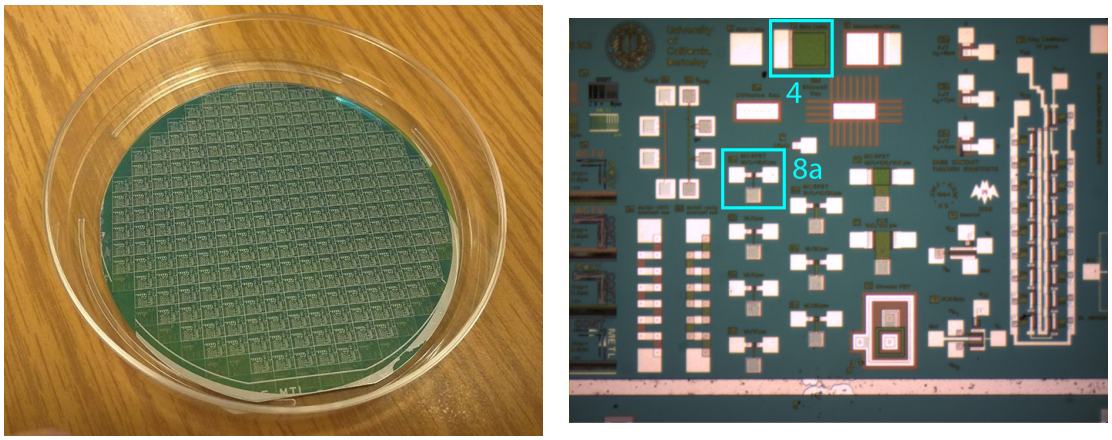
The above description highly suggests that the RBD mechanism is same as LBD. It is the combination of stress voltage and tunneling current that causes structural defects and finally leads to a conducting path. However, RBD happens in a much shorter time scale, such that device failure happens before a noticeable characteristic change. Also, after the

initial BD due to sustained high I-V stress, as well as the discharge of the gate capacitor, Joule heating causes destructive melting, from which it is impossible for the affected device to recover. In comparison, in the LBD cases, it is possible to monitor the slow, progressive change in the device characteristics, and it is even possible to repair the defective sites by re-annealing [94].

While EMI-induced circuit disruptions may put severe stress on the transistors, the breakdown progress may not be the same as under constant voltage due to the transient fluctuation of the stress. According to the oxide breakdown mechanisms, it is possible to evaluate the breakdown by looking at the state of the devices, such as the heat generation (temperature) and oxide field. Under disrupted conditions caused by EMI, the wear-out rate of LBD may fluctuate as the temperature and field may change in time. The device may also experience an extremely high field leading to an RBD. Thus, it is helpful to monitor the time-dependent device states under the desired interference, which may be different from the constant or high-frequency stress used in typical time-dependent dielectric breakdown studies [95, 96]. The state of the gate oxide could be used to estimate the LBD progress, based on the selected models. In the RBD case, an exception should be raised and the simulation should stop. It is then possible to report both types of oxide BD under the influence of EMI-induced transient stress.

5.2 Device Fabrication and Preliminary Tests

We fabricated the tested devices, as shown in Figure 5.1, at the FabLab at the University of Maryland NanoCenter.



(a) The whole wafer

(b) Microscopic view

Figure 5.1: Photos of a wafer fabricated using 0.5 µm process.
(a) A full view of the 3-inch wafer. The wafer contains 240 dies with 4 successful mask process steps.
(b) A microscope view of one die on the wafer. The two devices enclosed by light blue rectangles are MOSCAP (labeled "4") and MOSFET (labeled "8a"). They are the tested devices in this section.

A 3-inch Si wafer with p-type doping of $8 \times 10^{14}\,\mathrm{cm}^{-3}$ was used. Individual n-type MOSFETs of various dimensions were available on more than 200 dies. A self-alignment process with contact lithography was used. The mask channel length ranged from 4 µm to 20 µm. The gate oxide was thermally grown in dry O_2. The nominal oxide thickness was 80 nm. The n-type gate polysilicon was deposited using the LPCVD process. Spin-on phosphosilicate glass was deposited to form source and drain regions, followed by drive-in, isolation oxide growth, and dopant annealing. Aluminum was evaporated to create metal contacts.

Despite the large dimensions of our devices and thick oxide layers, they still fit for our purpose, which was to extract the gate dielectric breakdown (BD) condition and apply it to scaled-down devices with thinner gate oxide layers. Since the dielectric BD

largely depends on the E field ($E_{OX} = V_{OX}/t_{OX}$) rather than the gate stress voltage ($V_{OX} \approx V_{GS} - V_{TH}$) or gate oxide thickness (t_{OX}), for thinner oxides, the BD threshold condition for applied gate voltage (V_{GS}) can be adjusted accordingly while the ratio V_{OX}/t_{OX} leading to BD remains unchanged.

We carried out preliminary I-V measurements and stress tests on these fabricated devices on a probe station at room temperature with a signal source (Tektronix AFG3022), an oscilloscope (Tektronix TDS2014) and several DC supplies. Figure 5.2 is a photo of the test bench.

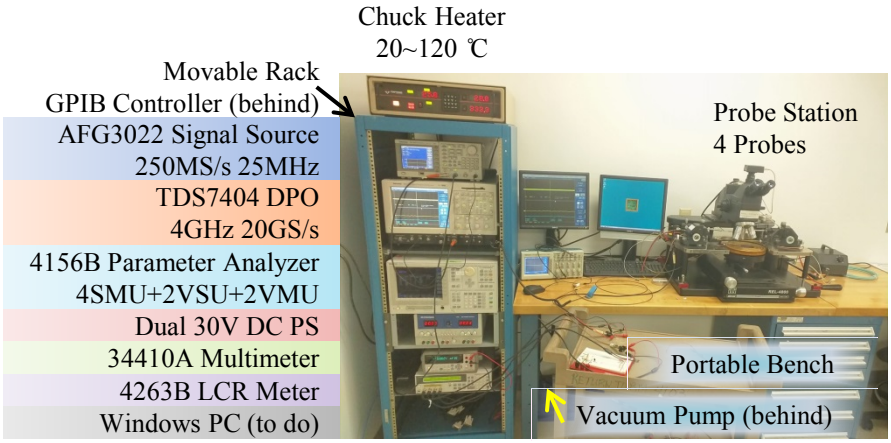
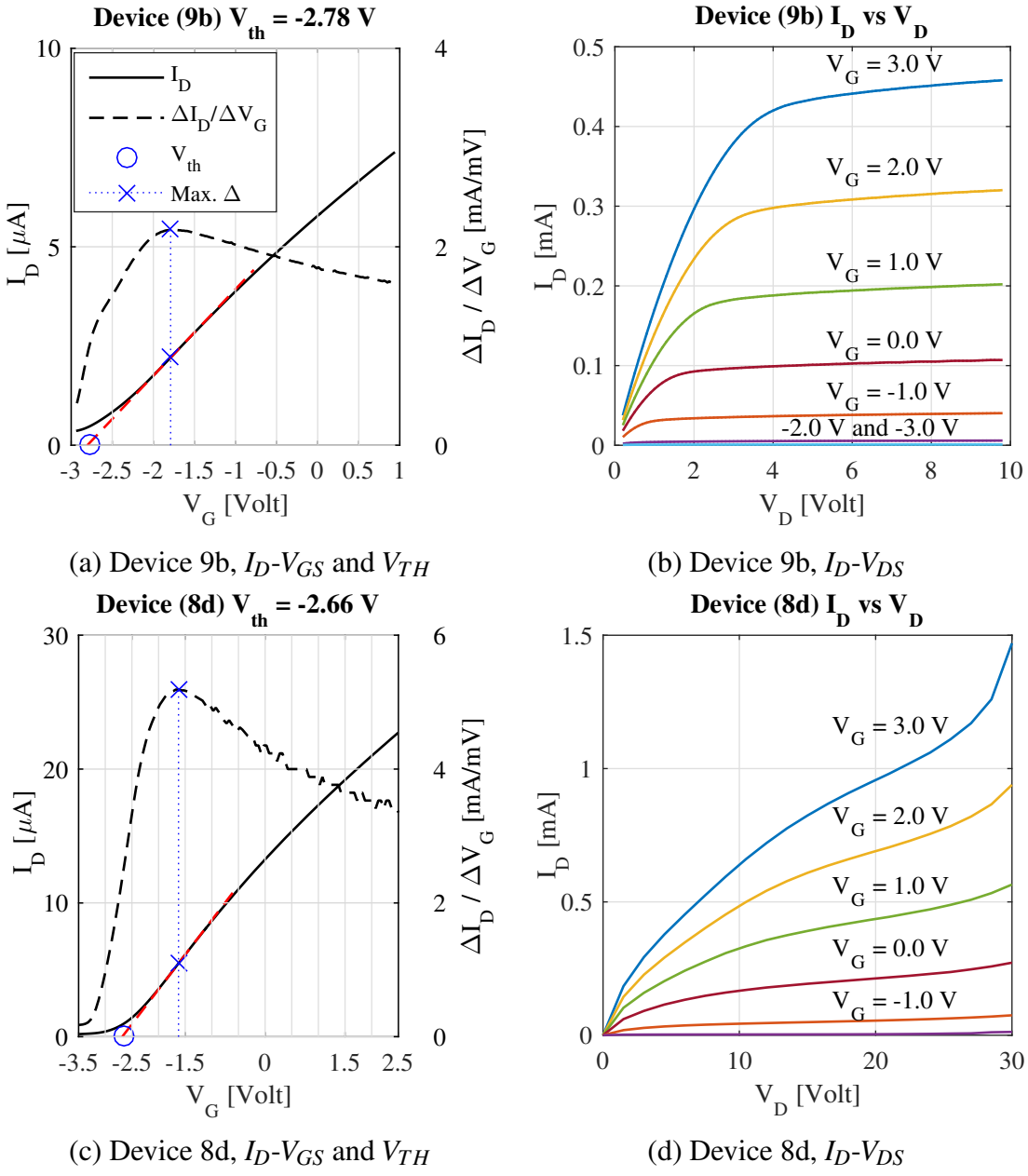


Figure 5.2: A photo of the experimental test bench that we developed, annotated.

First, the device I-V characteristics were measured. Examples of the I_D-V_{GS} and I_D-V_{DS} curves are provided in Figure 5.3. The threshold voltage was extracted with ELR method [78]. It ranged from $-2.6\,\mathrm{V}$ to $-2.9\,\mathrm{V}$. Clearly, it was much lower than expected ($V_{TH} \approx +0.1\,\mathrm{V}$ from the textbook formula). We believe it was due to positive charges in the gate oxide, introduced in the oxide growth step. In device 8d ($W \times L = 15 \times 10\,\mu\mathrm{m}^2$), severe current increases were observed under $V_{DS} = 25$–$30\,\mathrm{V}$. Considering the relatively low doping concentration in the substrate, it is highly possible that under such high reverse bias, the space charge region of the drain-body junction reaches the source-body junction, and consequently forward biases the latter. Effectively, the source-body barrier is lowered, and additional current parallel to the inversion channel appears. This is often characterized

as drain-induced barrier lowering (DIBL) in terms of threshold voltage shift (ΔV_{TH}) per increment of drain-source voltage (ΔV_{DS}). Meanwhile, impact ionization-induced avalanche current is also possible, similar to the measurement results in Section 3.1.



(a) Device 9b, I_D-V_{GS} and V_{TH}

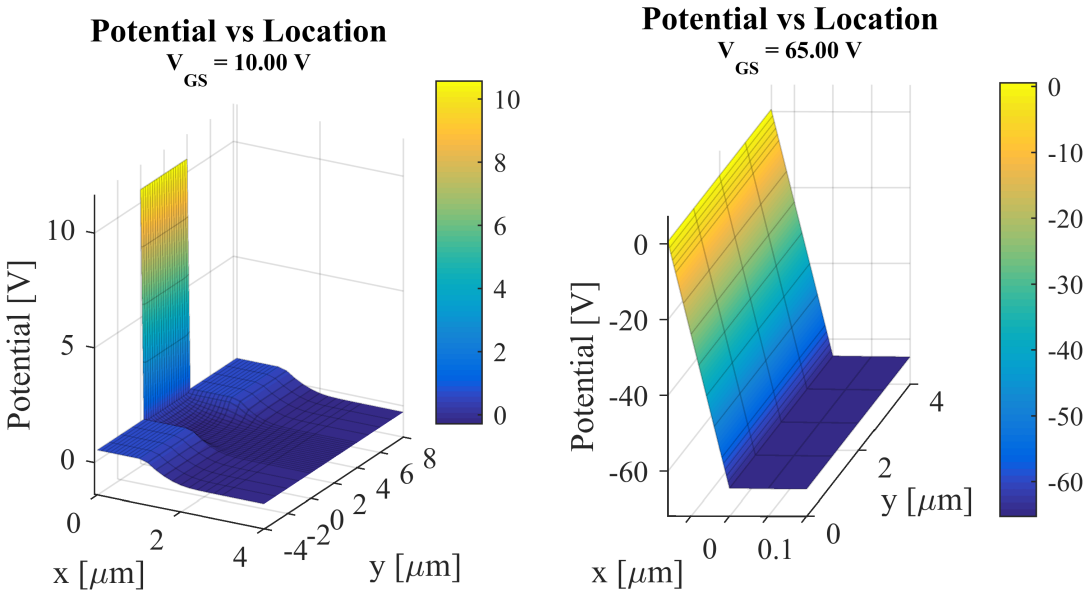
(b) Device 9b, I_D-V_{DS}

(c) Device 8d, I_D-V_{GS} and V_{TH}

(d) Device 8d, I_D-V_{DS}

Figure 5.3: Example of measured threshold voltage and characteristic I-V curves of our fabricated N-MOSFET devices.
(a) and (b) $W \times L = 15 \times 20\,\mu\text{m}^2$. $V_{TH} = -2.78\,\text{V}$.
(c) and (d) $W \times L = 15 \times 10\,\mu\text{m}^2$. $V_{TH} = -2.66\,\text{V}$. Under high gate and drain voltages, the drain current increased drastically, due to channel impact ionization.
Moving average was applied to all current data to reduce error in V_{TH} and visual fluctuation.

The devices were simulated, with one example shown in Figure 5.4, in order to measure the oxide field under various bias conditions. The device simulation did not include oxide breakdown mechanisms, so although the voltage might have reached a value that could cause BD, it was not reported by the program.



(a) MOSFET, $V_{GS} = 20\,\mathrm{V}$ (b) MOSCAP, $V_{GS} = 65\,\mathrm{V}$

Figure 5.4: Example of simulated spatial potential distribution in the devices we have fabricated.
(a) MOSFET, $W \times L = 15 \times 4\,\mu\mathrm{m}^2$. The source and drain of the MOSFET are in $-4\,\mu\mathrm{m} < y < 0$ and $4\,\mu\mathrm{m} < y < 8\,\mu\mathrm{m}$, respectively. Their terminals are left floating in the simulation and experiment; only the gate contact and substrate are connected.
(b) MOSCAP, actual size $W \times L = 200 \times 200\,\mu\mathrm{m}^2$, simulated size $L = 4\,\mu\mathrm{m}$.
The gate oxide ($T_{OX} = 80\,\mathrm{nm}$) is in $x < 0$. The substrate (body) terminals are at $x = 80\,\mu\mathrm{m}$ which is omitted in the figures.
The device simulation did not include oxide breakdown mechanisms, so although the voltage might have reached a value that could cause BD, it was not reported by the program.

5.3 DC RBD Test

We tested the devices under stress conditions. Rapid gate oxide BD was observed. In the first setup, shown in Figure 5.5, a DC voltage was added onto the gate polysilicon electrode. The wafer was loaded on a conducting chuck, which grounded the body of the device. By using manual control over the DC power supply, a range of stress voltages was swept, while the voltmeter readings were monitored.

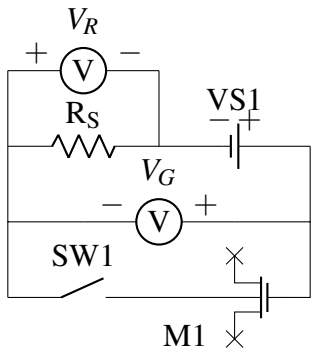


Figure 5.5: Test circuit of rapid oxide BD: a simple DC setup. The switch (SW1) protected the oxide of the N-MOSFET (M1) from being accidentally stressed before the setup was confirmed, or from thermal runaway after the oxide being broken and becoming conductive. The drain and source terminals of M1 were left non-connected.

The detailed steps were outlined in Figure 5.6. In each experiment, the gate oxide under test was first confirmed to be in its fresh state, by simply probing its conductivity under 20 V. A "good" oxide should be insulating, making the voltage divider circuit an open circuit and giving an ideal voltage measurement $V_G = V_{\mathrm{VS1}} = 20\,\mathrm{V}$ and $V_R = 0$. The stress voltage was then set, starting from 50 V (inducing a field of about $6\,\mathrm{MV/cm}$ in the oxide) and was increased by 1 V in each sweep step. In each sweep step, the time since the closing of the switch was recorded. The voltmeters were continually monitored. If oxide BD occurred, the oxide would become conductive, thus the meter readings would obviously and significantly change beyond the measurement noise. The stress was stopped when BD occurred, and the last stress condition was recorded, with the post-BD oxide resistance being measured by the voltage divider; otherwise, if 10 seconds had elapsed

without a BD event (oxide was still insulating), the stress voltage was increased and the test was repeated. A switch was used to protect the oxide from being accidentally stressed before the setup was confirmed, or from thermal runaway after the oxide being broken and becoming conductive.

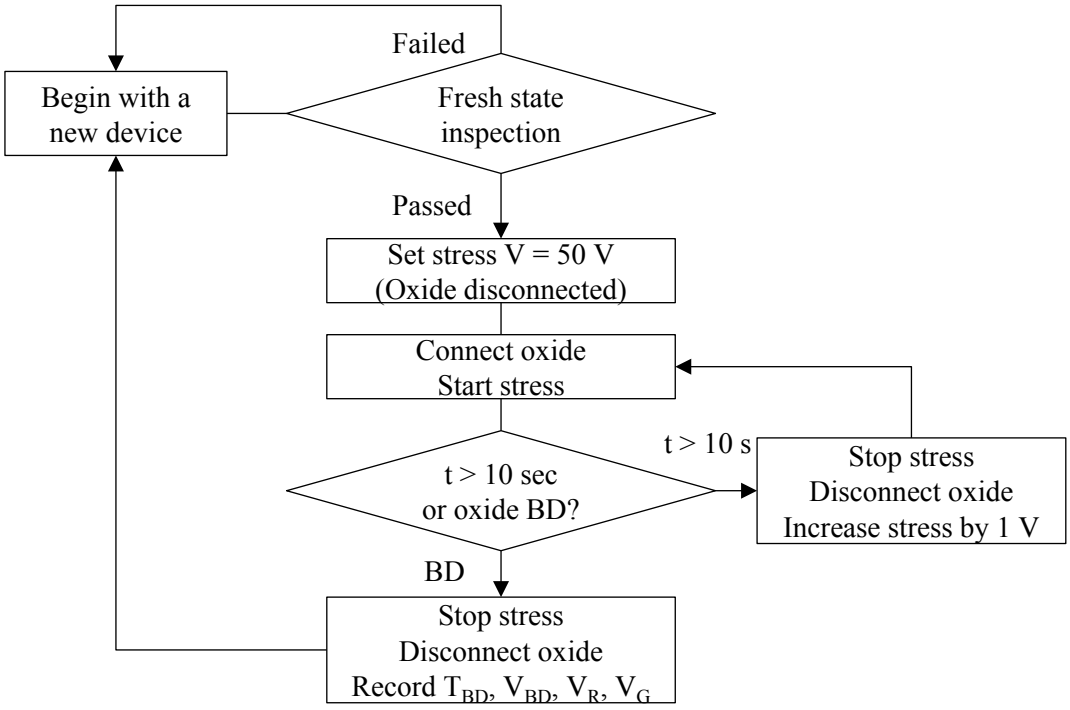


Figure 5.6: Flowchart of a simple DC stress test with our fabricated devices.
T_{BD}, V_{BD}, V_R and V_G stand for the time from closing the switch in Figure until the RBD event, the stress voltage before RBD, and voltmeter readings across the serial resistor and across the gate oxide, respectively.

We defined and measured the RBD condition variables as follow:

- The E-field during onset of RBD events (E_{BD}) was the device simulation result corresponding to the applied voltage V_{BD} when the onset of RBD was observed; an example simulation was shown in Figure 5.4. We increased the stress voltage by 1 V increments, starting from zero. At each voltage step, the oxide was stressed for 10 seconds. In the last stress step, the gate was stressed at V_{BD} for less than 10 seconds until the final RBD event. Therefore, E_{BD} represented the maximum field ever applied since the

beginning of test on an untested, fresh-state oxide and until the onset of RBD event. (Therefore, the oxide was not broken under 1 V less than V_{BD}.)

- The DC oxide resistance after BD (R_G) was measured by using the voltage divider circuit $R_G = V_G/V_R \cdot R_S$, where the power supply was adjusted to a lower value $V_{VS1} = 20\,\text{V}$. In this test, the external serial resistor $R_S = 100.4\,\text{k}\Omega$. The quantities V_R, V_G, and R_S are defined in the circuit schematic in Figure 5.5.

The DC stress test results were shown in Table 5.1, accompanied with peak oxide field in the last stress step from device simulation. We considered an RBD event to have occurred if a significant change in voltmeter readings were observed within 10 seconds of the stress for one gate voltage. The exact time before RBD was not accurately measured, but obvious abrupt changes in the oxide resistivity during the test within the specified time frame (10 seconds). Further improvements to the experiment provided accurate timing, which will be described in the transient RBD test in Section 5.4.

Table 5.1: Rapid oxide BD result: a simple DC setup. Measured devices were gate capacitors, $W \times L = 200 \times 200\,\mu\text{m}^2$.

Die #	RBD Conditions		Post-BD Resistance		
	V_{BD} [V]	E_{BD} [MV/cm]	V_R [V]	V_G [V]	R_G [kΩ]
(4,6)	54	6.76	7.40	12.7	58.3
(3,6)	50	6.26	5.84	14.2	41.1
(-7,5)	57	7.13	4.90	8.90	55.1
(-6,5)	59	7.38	9.10	11.0	82.7
(-5,5)	64	8.01	7.52	12.7	59.2

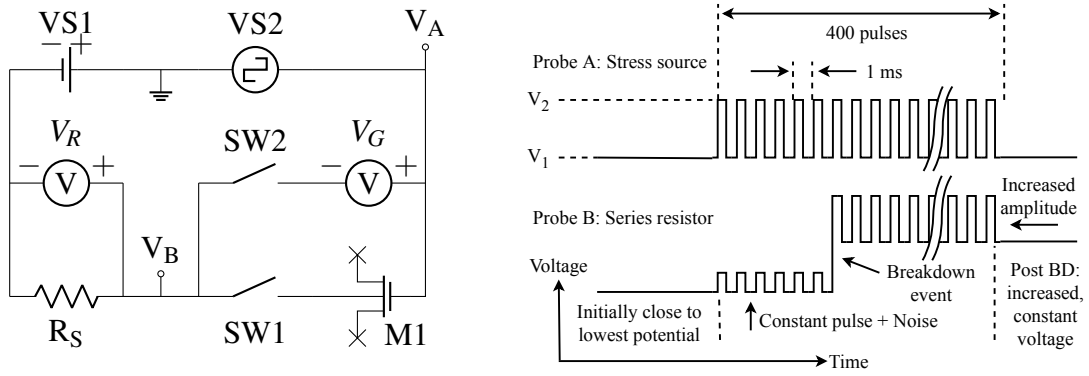
From the result, very little relation could be found between the voltage and field that brought to RBD, the time of the final stress before RBD, and the oxide resistance after RBD. However, we found consistent values of the BD field of 6 MV/cm to 8 MV/cm. Using this result as guidance, we improved our experiment configuration and included transient waveform stress with automated timing.

5.4 Transient RBD Test

In the next test setup, the idea of "stress-and-measure" from the previous DC test was continued, while transient stress signals were added, so we could accurately measure the time it took to an RBD event.

5.4.1 Test Setup and Workflow

The new test circuit and example measured waveforms are shown in Figure 5.7.



(a) Test circuit

(b) Signal source output

Figure 5.7: Our experimental setup for the transient RBD stress test.
(a) Test circuit. Quantities V_A and V_B are *measured* voltages.
(b) Illustrations of test voltage waveforms, measured with an oscilloscope. "Probe A" and "Probe B" correspond to V_A and V_B, respectively.

The newly added signal source VS2 provided transient waveforms (square waves) of up to 20 V of amplitude as a part of the stress signal. On the other hand, from the previous tests, we learned that the RBD events would occur under stress voltage in the range of 50 V to 70 V (which translates to a BD field of 6 MV/cm to 8 MV/cm). The DC power supply[18] VS1 provided the additional stress voltage (40 V or 50 V following the test workflow, to be described shortly).

[18] Since a single DC power supply in our lab could provide up to 30 V, in reality, VS1 was a combination of three PS's in series connection. Using a redundant unit, the DC output was switched between two levels on demand.

An oscilloscope was used to measure and observe the voltage waveforms at V_A and V_B during the stress test, so that the onset RBD events were captured and timed. Hence, besides the two RBD conditions defined previously in Section 5.3, we have one more variable measured in this improved test. The three RBD condition variables are:

- The time to BD (T_{BD}) was the total stress time when the signal source (VS2) was generating the last and highest amplitude V_{BD}. The stress voltage at the onset of RBD event V_{BD} was the sum of voltages provided by the constant DC source VS1 (40 V or 50 V depending on the test flow described in Figure 5.8) and the signal source VS2 (the high level V_2 in Figure 5.7b).
- The E-field during onset of RBD events (E_{BD}) was defined previously. It corresponded to the highest applied voltage before BD (V_{BD}).
- The DC oxide resistance after BD (R_G) was defined previously.

Furthermore, to avoid faulty short connections between the reference potential of different devices, all active devices were connected to a common ground – the DC power supply (VS1), the signal generator (VS2), and the oscilloscope. As a result, signals V_A and V_B in Figure 5.7b ranged from -50 V to $+20$ V, and the actual potential difference across the MOS device under test is $V_A - V_B$. An additional switch SW2 isolated the voltmeter (V_G) from the rest of the circuit during the stress and measurement phase, which could be affected by the meter's internal resistance.

Probe A signal (V_A) was the signal generator output between V_1 and V_2 as in Figure 5.7b. At rest, the source stayed at low output ($V_A = V_1 = 0$ V before RBD). After a triggering event, it generated square wave pulses of a limited length $t_{MAX} = 400$ ms or 400 cycles. After t_{MAX}, it then returned to the low output and stayed until the next triggering event. The square wave had high level $V_A = V_2$, period 1 ms and duty cycle 50%. V_2 was increased from 1 V to 10 V or 20 V (following the test workflow, to be described shortly) by increment steps of 1 V.

Probe B signal (V_B) was very close to the DC power supply (VS1) output before the RBD event. When the signal source (VS2) was triggered, voltage pulses were present. Although the oxide was not conductive prior to the RBD event, due to the capacitive displacement current, a very small square wave could be observed in V_B, but the amplitude was small and comparable to the background noise.

When an RBD event occurred, the square wave amplitude suddenly increased or "jumped" to a much higher level until the transient stress finished in t_{MAX}. In the end, V_B stayed at a much higher DC voltage than VS1's output, because the broken oxide was conducting DC current. Hence, the voltage condition at BD was $V_{BD} = V_1 - V_B$, where V_1 and V_B were the last set values according to the workflow described as follows.

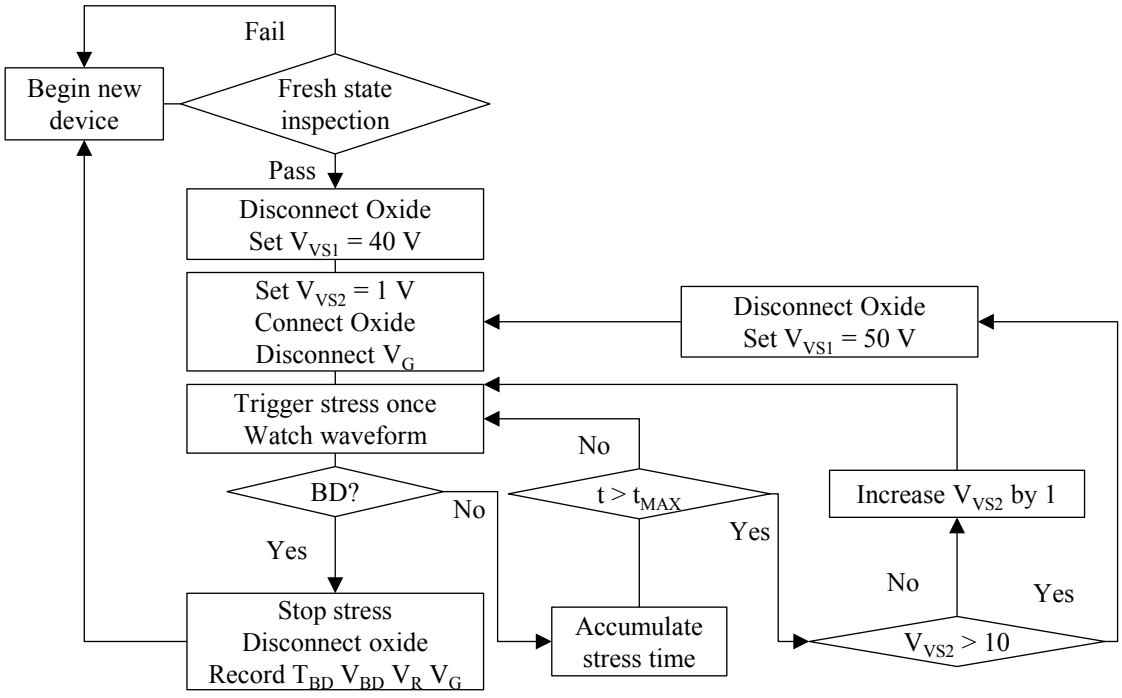


Figure 5.8: A flowchart of the transient stress test we performed on our fabricated devices. The quantity names have the same meanings as in Figure 5.6. The oxide always broke down before high level reached 70 V, so no infinite loop was created.

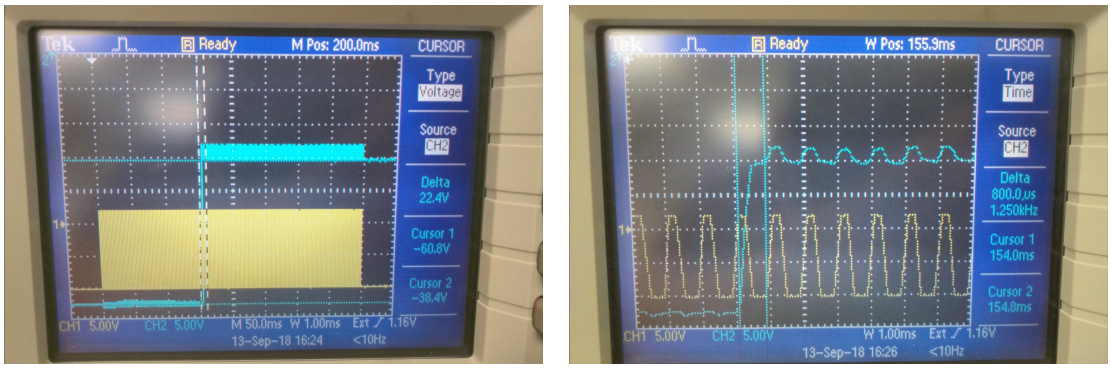
The augmented test procedure was outlined by the flowchart in Figure 5.8. The DC voltage source VS1 (see Figure 5.7) applied the non-destructive 40 V to the oxide, while the signal source VS2 stayed at zero. This DC bias from VS1 was applied for the duration

of the experiment, and it was not considered to cause RBD since the corresponding oxide field was below the threshold range, according to our DC test results.

The signal source VS2 generated a constant voltage $V_A = V_1 = 0\,\mathrm{V}$ (see Figure 5.7) until it was triggered, after which it output a limited number of square waves (pulses) of period 1 ms, duty cycle 50% for $t_{MAX} = 400\,\mathrm{ms}$ (400 cycles). Starting from $V_A = V_2 = 1\,\mathrm{V}$, the amplitude of the pulse waves was increased by 1 V at each step. For each voltage step, the signal source VS2 and the oscilloscope were triggered synchronously, and the finite-time transient stress signal was applied; this was one "run". If the oxide did not break down, the transient stress amplitude V_2 was increased by 1 V, and the "runs" were repeated until RBD was observed. Effectively, the timer for T_{BD} was reset before each "run".

If the transient voltage amplitude reached $V_2 = 10\,\mathrm{V}$, the DC voltage supply was increased from 40 V to 50 V, V_2 was reset to 0 V, and the test continued. The oxide always broke before the total applied voltage $V_2 - V_B$ reached 70 V.

The oscilloscope was set at a high resolution and long record length, so that it captured all "burst" pulses with enough samples to describe the individual pulses. One example was provided in Figure 5.9, including an RBD event.



(a) A whole "run" (b) Zoom-in view

Figure 5.9: An example of the transient stress measurement on our fabricated device. The device under test was a MOSCAP, $W \times L = 200 \times 200\,\mu m^2$. Channel 1 (yellow) was Probe A (V_A). Channel 2 (light blue) was Probe B (V_B).
(a) A whole "run" was shown. The trigger event occurred at the time zero indicated by the time reference arrow (upper left corner).
(b) Zoom-in view around the RBD event, where the time to BD was accurately read out ($T_{BD} = 154.0\,ms$).

The stress duration captured in one "run", as shown in Figure 5.9a, lasted $t_{MAX} = 400\,ms$ following a manually input trigger signal. The pulsed stress (1 kHz, $-10.2\,V$ to $2.00\,V$) lasted for 400 ms, including 400 cycles. Initially, Probe B signal stayed at $V_B = -60.8\,V$. When transient stress was present, there were small ripples resulting from the oxide capacitor current. At time of approximately 150 ms from the time reference point, there was an obvious and abrupt change in the probe voltage V_B (see Figure 5.7a), indicating an RBD event. This obvious event determined the end of the test, and the time from the triggering event to the onset of RBD was the time to BD (T_{BD}).

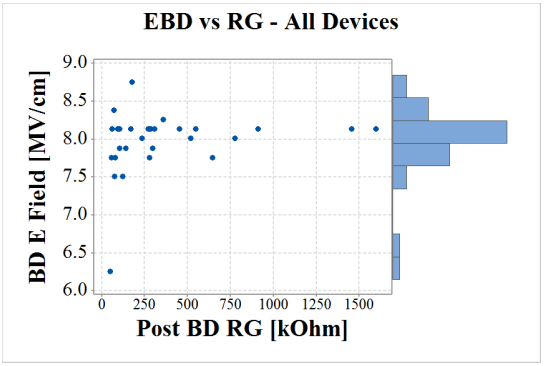
Then, we examine the test waveforms more carefully. A close-up view of the waveform when the RBD occurred was shown in Figure 5.9b. After the BD event, the square wave in V_B had a low level of $-38.4\,V$ and a high level of $-36.0\,V$. From this zoom-in view, we could read the accurate time to BD as $T_{BD} = 154.0\,ms$. It was clear that the RBD

event took place within less than one stress pulse of 1 ms, and there was no observable gradual changes before or after the onset of BD event.

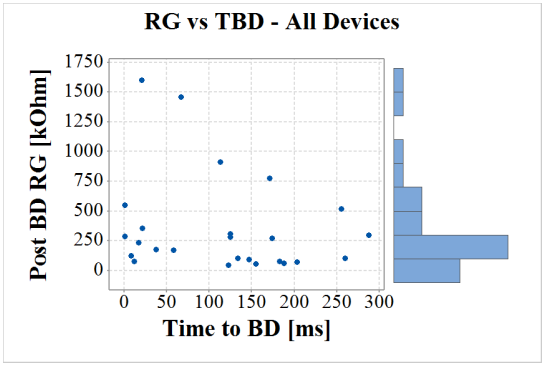
A measurement using the voltage divider circuit showed the oxide in this example had a DC resistance of 126 kΩ after RBD, compared to virtually infinity before the stressing, as no measurable current was present. Its small signal resistance was 408 kΩ. This was unexpected since the measured resistance should be smaller at higher frequencies, because of the gate capacitance. One possible reason was our test bench contained serial inductance, adding to the measured high-frequency impedance.

5.4.2 Test Results and Analyses

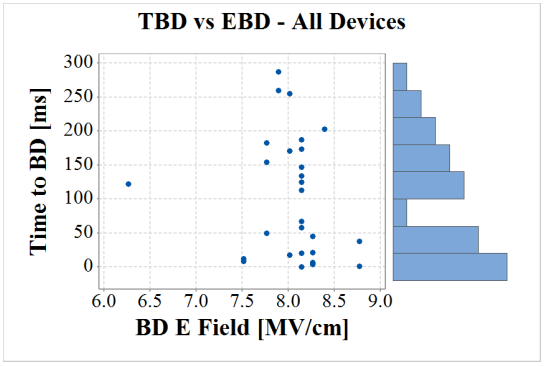
We tested thirty-five (35) devices of two sizes. We collected the experimental data and analyzed it using the statistical software Minitab [97]. The results were plotted and shown in Figures 5.10 and 5.11.



(a) E_{BD} vs R_G



(b) R_G vs T_{BD}



(c) T_{BD} vs E_{BD}

Figure 5.10: Transient RBD stress test data on our fabricated devices, shown in scatter plots. The devices under test were MOSCAP's, $W \times L = 200 \times 200\,\mu m^2$, and MOSFET's, $W \times L = 15 \times 4\,\mu m^2$. The total sample size was 35. RBD field E_{BD}, time to RBD T_{BD}, and post-BD gate resistance R_G were plotted in pairs of two.
E_{BD} was the device simulation result that corresponded to the high level of the transient pulses (V_2 in Figure 5.7).
The accumulated stress time T_{BD} included all "runs" with the last stress voltage before RBD. It was the total time when the signal source was generating transient stress signals (square waves with 50% duty cycles), including whole duty cycles.

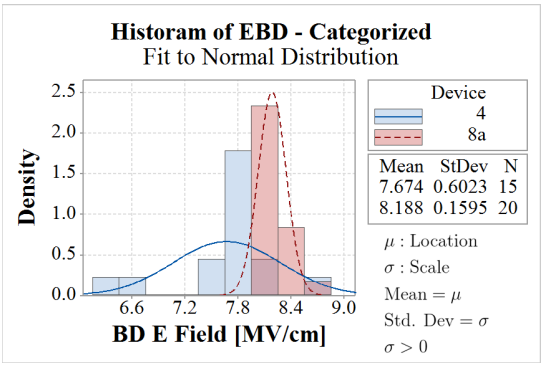
In Figure 5.10, the RBD measurement sample points of all devices were mixed, independent of their sizes. Three variables were plotted in x-y plots in an attempt to find any relationship between each pair of two. However, no strong dependency was observed,

and instead, the samples seemed to loosely follow three independent distributions with respect to the three observed variables (distribution definitions are in Table 5.11d):

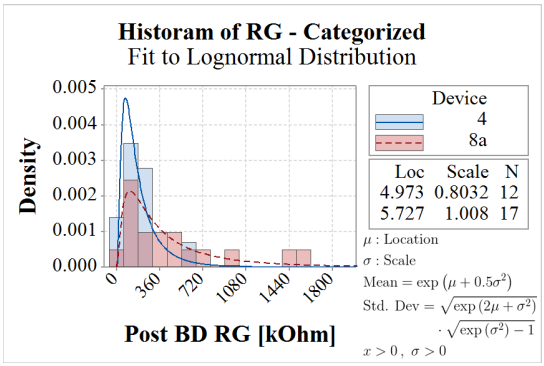
- An exponential distribution for the time to RBD (T_{BD}),
- A normal distribution for the E-field during onset of RBD events (E_{BD}), and
- A log-normal distribution for the resistance after RBD events (R_G).

The marginal distribution histogram plots were drawn on the side of the x-y plots in Figure 5.10. The results indicated that more devices tended to break down sooner (or in fewer stress cycles), and the measurement error was a main contributing factor to the spread in the data.

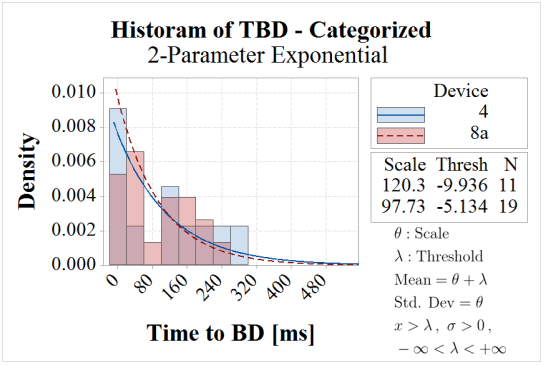
In Figure 5.11, the sample histograms were drawn, categorized by device types, and the analytical probability density functions were listed.



(a) E_{BD} and normal distribution



(b) R_G and lognormal distribution



(c) T_{BD} and exponential distribution

Distribution	PDF $f(x)$
Normal	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
Lognormal	$\frac{1}{\sqrt{2\pi}\sigma x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$
Exponential	$\frac{1}{\theta}\exp\left(-\frac{x-\lambda}{\theta}\right)$

(d) Probability Density Functions. The dummy variable "x" is explained as below.

Figure 5.11: Scatter plots of the transient stress RBD test data on our fabricated devices, with categories of two types of devices under test, MOSCAP's (labeled "4"), $W \times L = 200 \times 200\,\mu m^2$, and MOSFET's (labeled "8a"), $W \times L = 15 \times 4\,\mu m^2$.
(a) The RBD field E_{BD} was fitted to normal distribution.
(b) The post-BD gate resistance R_G was fitted to log-normal distribution.
(c) The time to BD T_{BD} was fitted to exponential distribution.
In the histograms, "Density" = (Normalized frequency) / (Bin width). The sum of all bin heights within one group of devices, multiplied by the uniform bin width, equals to 1. The probability density functions of the three distributions were listed in Table (d).

Most devices broke down under a field of around 7 to 8 MV/cm, and the average resistance values after BD were 199 kΩ for the larger devices and 510 kΩ for the smaller devices. As stated before, more devices tended to break down earlier in time, indicating that the RBD events were more likely to happen instantaneously. Also, it could be seen that there was little difference in T_{BD} among the two groups. Among our thirty (30) tested devices with valid T_{BD} measurements, 50 % broke down within $T_{BD} < 100$ ms, and 87 % broke down within $T_{BD} < 200$ ms.

It is interesting to notice that when the area of the gate oxide layer increased by about 600 times, the difference of the BD time between the two groups and the reduction in the resistance of the broken oxides right after RBD events were small. Both the MOS capacitors and MOSFETs on our fabricated wafer had the same "gate" oxide thickness, which was fully covered by the polysilicon layer as the gate terminal. On the other hand, the average BD field had about 5% difference between the two groups, with the larger devices having lower magnitudes. This could be a result of the larger area having more defects or impurities introduced in fabrication, but the evidence was weak considering the limited sample size.

Finally, long-term post-RBD stress was performed, and the slow decrease in the (broken) oxide resistance was observed, with data shown in Table 5.2, and one example of the captured signal shown in Figure 5.12. In each post-RBD test, the stress voltage was lowered to 30 V, DC only. After 100 s of stress, there was a 3% to 11% decrease in R_G. This indicated that the gate current after RBD increased and the damage got worse, which could be because of the growth of the initial BD sites.

Table 5.2: Measurement results of post-RBD long-term (broken) gate resistance decrease under constant, low-level stress (30 V). Tested devices were gate capacitors (labeled "4"), $W \times L = 200 \times 200\,\mu m^2$ and MOSFET's (labeled "8a"), $W \times L = 15 \times 4\,\mu m^2$.
The peak magnitude of the oxide electric field was extracted from the device simulation mentioned earlier, with an example shown in Figure 5.4.
V_B was the voltage measured in Figure 5.12. A steady noise of 200 mV was present. Shown in the table was the average of the min and max values at $t = 0\,s$ and $t = 100\,s$.
The gate resistance R_G was calculated using the circuit in Figure 5.7a, with SW2 open. $R_G = V_G / (V_R/R_S + V_B/R_P)$, where the serial resistance $R_S = 100.4\,k\Omega$, oscilloscope probe internal resistance $R_P = 1\,M\Omega$, DC power supplies $V_{VS1} = -33.2\,V$, $V_A = 0\,V$, $V_G = V_A - V_B$ and $V_R = V_B - V_{VS1}$.

		Right after RBD		After 100 s stress		
Die / Dev #	E_{BD} [MV/cm][1]	V_B [V]	R_G [kΩ]	V_B [V]	R_G [kΩ]	ΔR_G
(-5,6)/8a	8.26	-27.9	719	-27.7	654	-9.12%
(-4,6)/4	7.76	-23.4	185	-23.3	177	-4.35%
(-3,6)/4	8.63	-19.1	75.0	-18.9	72.0	-3.93%
(-3,6)/8a	8.51	-22.2	140	-21.6	123	-11.6%

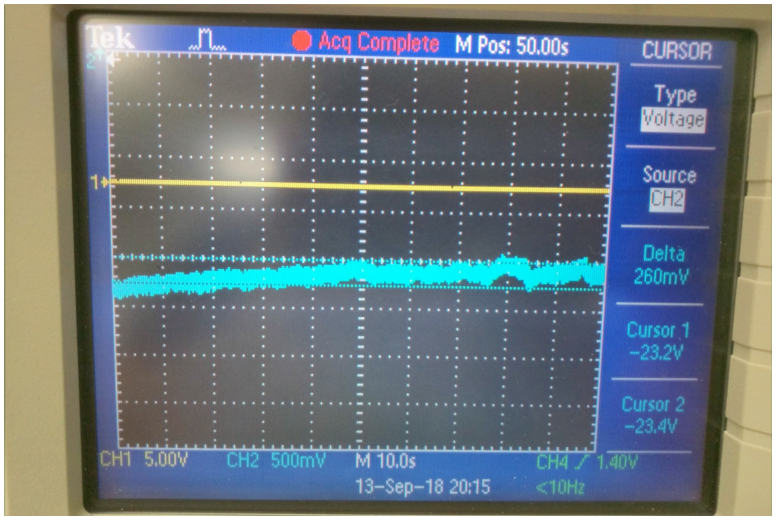


Figure 5.12: An example of the post-RBD long-term resistance change measurement on our fabricated device. The device under test was a MOSCAP, $W \times L = 200 \times 200\,\mu m^2$. It was stressed at a lower constant voltage 30 V for 100 s. V_B appeared as Channel 2 (light blue). Despite the $\sim 200\,mV$ noise, a steady change in the signal could be seen, indicating a decrease in the oxide resistance over time.

We summarize our RBD experiments with the following conclusions. We believe that one of the threshold conditions for RBD is the highest E field the oxide has experienced. Most devices break down when the field is around $E_{BD} = 7\text{–}8\,\mathrm{MV/cm}$. For the larger oxides ($200 \times 200\mu\mathrm{m}^2$), the average threshold is $7.674\,\mathrm{MV/cm}$, and $8.188\,\mathrm{MV/cm}$ for the smaller oxides ($15 \times 4\mu\mathrm{m}^2$). The RBD events occur within a very short time window after the stress has been applied and maintained, i.e., $T_{BD} < 200\,\mathrm{ms}$. Most devices, on average, have post-BD resistance of $R_{BD} = 199\,\mathrm{k}\Omega$ (larger size) or $R_{BD} = 510\,\mathrm{k}\Omega$ (smaller size) depending on the size of the devices. The continued stress after an initial RBD event can substantially and progressively decrease the resistance R_{BD}. There are weak to minimal relations between the above three measured variables, despite the larger-size group generally appears more vulnerable.

In other words, the severeness of RBD weakly depends on the oxide area, suggesting agreement with the BD mechanisms reviewed in Section 5.1. The initial RBD event occurs relatively fast. After that, due to the new presence of electrons in localized defect states, the broken site grows and spreads, resulting in more conductive paths. The possibility of having both the initial and following broken sites is generally related to the oxide defect density and the oxide area. This agrees with our experimental findings that the device group with larger areas has a slightly lower E_{BD} threshold and a slightly smaller R_{BD}.

Chapter 6: EMI-Induced Hard Failure Vulnerabilities: Nano-Scale FinFET Device-Level Simulation and Empirical Oxide Breakdown Circuit Model

In this chapter, we project our knowledge of RBD conditions in planar MOSFETs with SiO_2 gates to FinFETs with ultra-thin SiO_2 gate dielectrics. In Section 6.1, we briefly introduce FinFETs by looking at a figure of merit, the natural length, which is related to improvements in the short channel effects. Next in Section **??**, we investigate the electrical perspective of a FinFET structure with the most fundamental features. Device-level simulations (Section 6.2) provide characteristic behaviors such as electron concentration, internal electric field, and terminal I-V relationships. A quantum-corrected system of equations under quasi-equilibrium conditions is solved, and the results are compared to the classical device-level solutions (Section 6.3), demonstrating the characteristic changes due to the mesoscopic quantum well in the nano-scale devices.

Extreme gate oxide field and rapid oxide BD can be caused by terminal voltage disruptions when EMI is present. Therefore, external stress may be transient but catastrophic. We extract useful information from the simulation data concerning the Rapid oxide BD (RBD) in Section 6.3.4. Thus, the Hard Failure vulnerability due to RBD under EMI conditions can be modeled. In Section 6.4, a SPICE-compatible circuit-level model is developed using empirical data and expressions.

6.1 Introduction: Silicon FinFET Devices

FinFETs are one of the most widely used devices at present [52, 54]. A FinFET can have multiple (usually two or three) gates, protruding into the third dimension (height or depth). An apparent benefit of having multiple gates in different dimensions is that they can provide much more channel width (and therefore more terminal current) per unit area of the silicon die, compared to a traditional planar device. The adaptation of FinFETs is widely seen as we continue to advance the progress of area reduction, density increase and performance improvement [52].

The FinFET can also provide better gate control, thanks to its unique structure. Generally speaking, when a device's channel becomes shorter and shorter (especially when $< 100\,\text{nm}$), several problems collectively known as the *short-channel effects* become more and more significant. For example, the drain bias may cause unwanted, gate-uncontrolled current increase, when the device is in the OFF state (drain-induced barrier lowering or DIBL [98]). The same problem may happen in the ON state (this phenomenon has a confusing name, which is the short-channel effect, or SCE [99]; it is a different interpretation to channel length modulation). The 3D structure can provide very thin body regions which can suppress the short-channel effects by increasing the gate control.

The gate control effectiveness of various three-dimensional, multi-gate MOSFET structures, including FinFET, is described by the *natural length*, a geometry-dependent figure of merit [54].

$$\text{Natural Length: } \lambda = \sqrt{\frac{1}{N_G}\frac{\varepsilon_{Si}}{\varepsilon_{OX}} W_{BODY}\, t_{OX}} \tag{6.1}$$

The natural length is related to the exponential decay rate of the body potential and the channel lateral field (induced by the drain bias). It estimates the minimum required channel

length to avoid excessive short-channel effects. It has several dependencies:

$$N_G = \textit{Effective Number of Gates}$$

$$W_{BODY} = \textit{Body thickness} \text{ or } \textit{Fin width}$$

$$\varepsilon_{OX} = \text{Gate oxide dielectric constant}$$

$$t_{OX} = \text{Gate oxide thickness}$$

To achieve shorter gate lengths while maintaining the gate control, one may reduce the gate oxide thickness t_{OX}, but a very thin oxide layer (less than 1.5 to 2 nm) suffers from significant tunneling current [89]. Using high-K materials to increase ε_{OX} unavoidably adds to the production cost.

Alternatively, one can increase the number of gates N_G or reduce the body (fin) width W_{BODY}, which is only available to devices with three-dimensional structures such as FinFETs, but not to planar devices. A FinFET can provide an effective N_G ranging from 2 to 4. The body width W_{BODY} is largely reduced compared to planar devices.

The FinFET devices simulated in the following studies are based on the bulk-tied, "Π"-shaped triple-gate structure[19]. It has an N_G close to π. The poly-silicon layers on the three gates are electrically connected to form a single gate.

Despite the benefits of higher device density and better gate control, FinFETs may still be vulnerable to EMI. As the channel becomes shorter and the gate oxide becomes thinner, a transient voltage disturbance may cause a higher lateral E field in the channel and a higher perpendicular E field in the gate oxide. The large field in the channel may still cause impact ionization, although ballistic transport is more likely in a shorter channel. The parasitic BJT structure still exists in FinFETs, suggesting the possibility of snapback [100]. Additionally, the high oxide field may cause permanent oxide breakdown or accelerated progressive degradation. Furthermore, the heat dissipation is generally more challenging, as the body is surrounded by the gate and isolation structures [101, 102]. According to

[19]The word "FinFET" used here and from now on particularly refers to the "Π-gate FET" in the cited work [54].

the thermochemical model explained in Section 5.1, the oxide breakdown process may accelerate under elevated temperature (Equation 5.1). Reliability issues related to snapback and dielectric breakdown have been reported [100, 103–107]. Under extreme heat conditions, the lattice structure may even be damaged. Therefore, applying the knowledge and methodologies on EMI-induced device vulnerabilities to FinFETs, as well as other advanced structures, has great importance for modern technologies.

6.2 FinFET Device-Level Electrical Simulation (2D)

In this section, a silicon FinFET with bulk-connected bodies (Π-FinFET [108]) are simulated to obtain their electrical behaviors (e.g., current-voltage relationships).

Critical geometric layout parameters are illustrated in Figure 6.1. The body is a protruding region from the substrate material, with the gate stack surrounding it. The body is connected to the substrate with the same type of material (silicon), which is the fundamental difference compared to the silicon-on-insulator (SOI) structure. The quantities describing the geometric structure and doping profile are listed in Table 6.1 and Table 6.2 [109, 110].

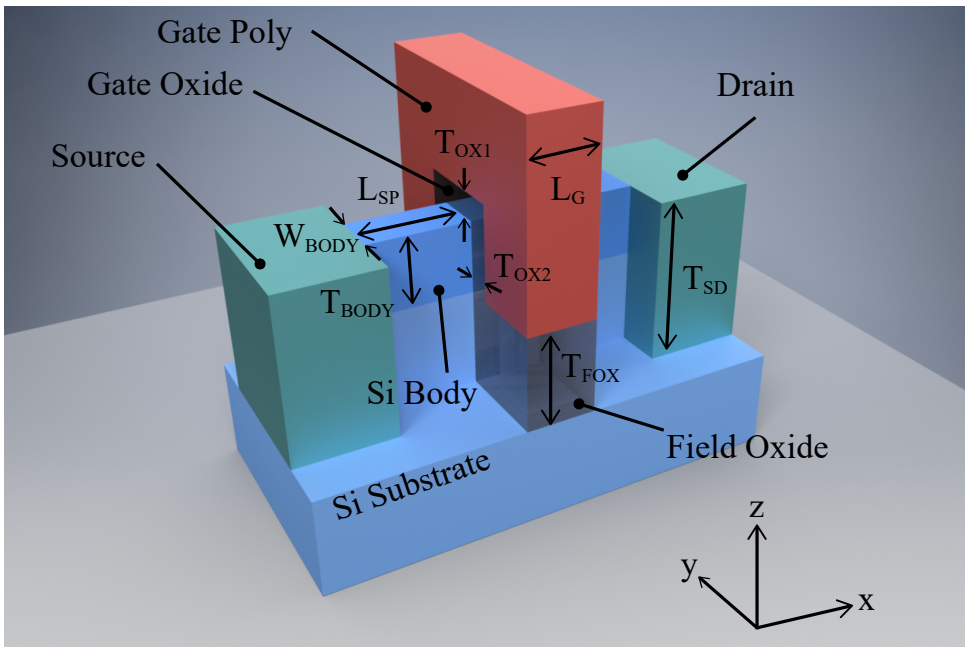


Figure 6.1: Illustration of a Π-FinFET structure in 3D and geometrical dimensions. The dimensions (listed in Table 6.1) are not shown in correct scale. The silicon body and substrate are of the same material and are physically connected. In the graph, different colors and transparency are used to differentiate the conceptual regions.

Table 6.1: Geometric dimensions of the simulated Π-FinFET, as illustrated in Figure 6.1

Critical dimensions		Value [nm]
L_G	Gate / channel length	25
L_{SP}	Side wall spacer length	50
$T_{OX1} = T_{OX2}$	Gate oxide thickness	3
T_{BODY}	Body (fin) height	34
W_{BODY}	Body fin width	15
Non-critical dimensions		
T_{FOX}	Field oxide thickness	100
T_{DS}	Fin / drain / source height	$T_{BODY} + T_{FOX}$

Table 6.2: Doping profile of the simulated Π-FinFET

Region (Dopant type)	Shape profile and concentration
Substrate (p)	Uniform, effectively $2.2 \times 10^{18}\,\mathrm{cm}^{-3}$
Drain and Source (n+)	Error function (Erfc), peak $1 \times 10^{20}\,\mathrm{cm}^{-3}$
Gate poly-Si contact (n++)	Uniform, $1 \times 10^{20}\,\mathrm{cm}^{-3}$

The bulk Π-FinFET is simulated using the 2D TCAD simulator Cider included in ngspice [62]. A "slice" is taken through the three-dimensional device to create a two-dimensional geometry to simulate. First, a top-down slice is made across the middle of the fin. The full simulation region and doping profile are shown in Figure 6.2. The body region is the box where $0 \leq x \leq L_G = 25\,\mathrm{nm}$ and $0 \leq y \leq W_{BODY} = 10\,\mathrm{nm}$. The gate oxide ($SiO_2$) appears as the gaps in $-3\,\mathrm{nm} \leq y \leq 0$ and $10\,\mathrm{nm} \leq y \leq 13\,\mathrm{nm}$, while the gate electrode appears as the boxes in $-6\,\mathrm{nm} \leq y \leq -3\,\mathrm{nm}$ and $13\,\mathrm{nm} \leq y \leq 16\,\mathrm{nm}$. The other gaps surrounding the gate electrodes are field oxides (SiO_2). In $-50\,\mathrm{nm} \leq x \leq 0$ and $25\,\mathrm{nm} \leq x \leq 75\,\mathrm{nm}$, the body extends through the sidewall spacer before expanding in the crosswise (y) direction and meeting the source and drain electrodes.

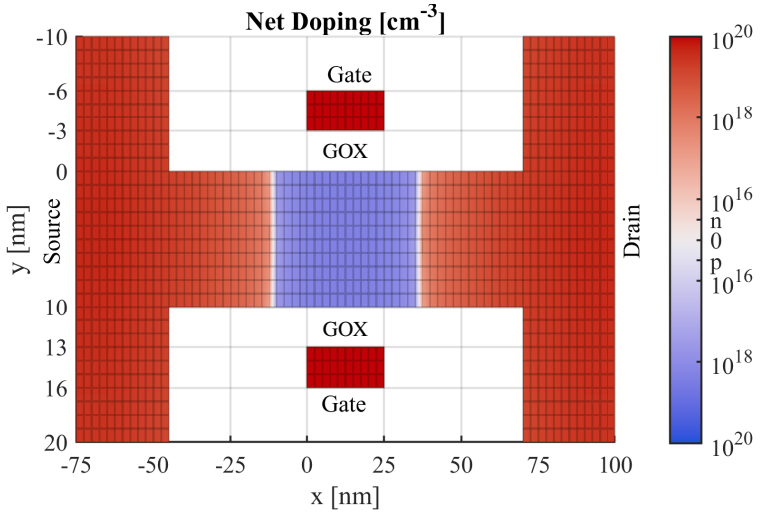


Figure 6.2: Doping Profile of the Simulated FinFET in top-down cross section view. The entire simulation region is shown. The color indicates doping concentration of n (red) and p (blue) types.
The empty regions are filled with dielectric (SiO_2) during simulation. The gate oxide ("GOX") is between $0 \leq x \leq 25\,\text{nm}$ and $-3 \leq y \leq 0\,\text{nm}$ for the upper side, and $10 \leq y \leq 13\,\text{nm}$ for the lower side.
The lines are visual guides revealing a part of the mesh grids.

The source and drain doping consists of two doses of n-type, $5 \times 10^{19}\,\text{cm}^{-3}$ in $-75\,\text{nm} \leq x \leq -65\,\text{nm}$ and $90\,\text{nm} \leq x \leq 100\,\text{nm}$ with an error function (erfc) grading with characteristic length 45 nm, and two doses of p-type, $5 \times 10^{18}\,\text{cm}^{-3}$ in $-75\,\text{nm} \leq x \leq -30\,\text{nm}$ and $55\,\text{nm} \leq x \leq 100\,\text{nm}$ with an erfc grading with characteristic length 55 nm. The entire substrate has a background p-type doping of $1 \times 10^{18}\,\text{cm}^{-3}$.

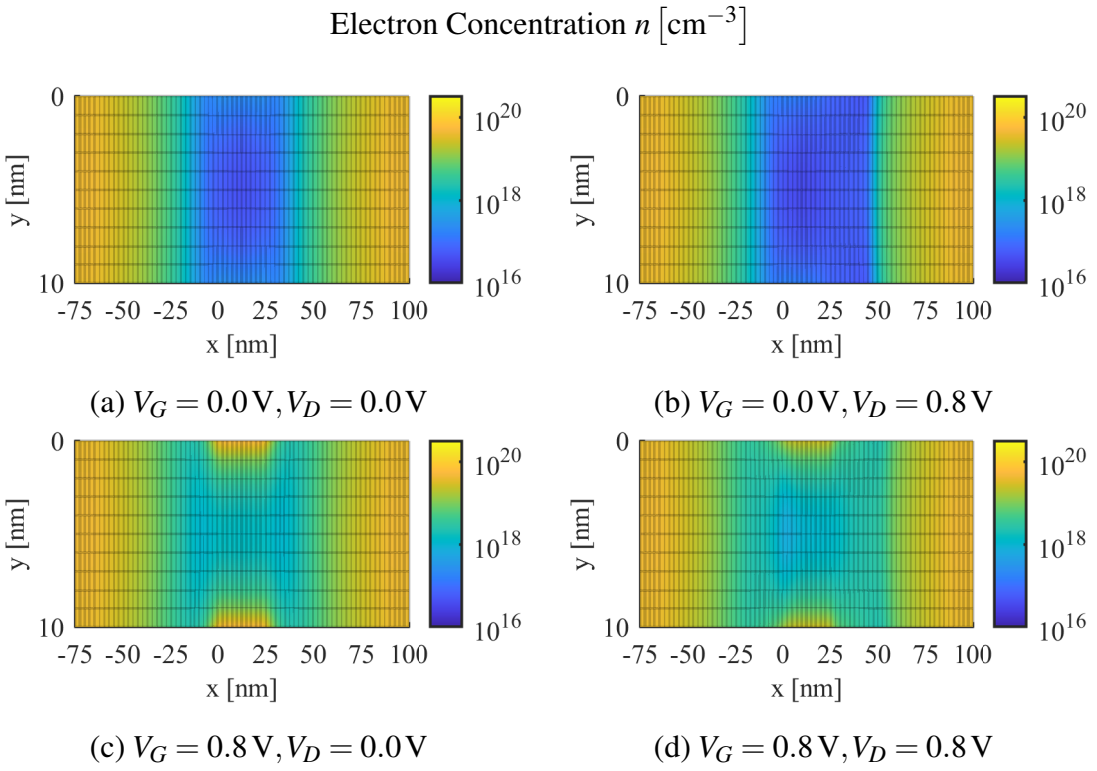
The junction space charge region size can be estimated by the textbook expression using depletion approximation and abrupt junction:

$$x_p = \sqrt{\frac{2\varepsilon_{Si}}{q}\frac{N_D}{N_A}\frac{1}{N_D+N_A}\phi_0} = 25\,\text{nm} \tag{6.2}$$

where $N_D = 1 \times 10^{20}\,\text{cm}^{-3}$, $N_A = 2.2 \times 10^{18}\,\text{cm}^{-3}$, and $\phi_0 = V_T \ln\frac{N_D N_A}{n_i^2} = 1.1\,\text{V}$. This distance is comparable to the gate, or the intended channel length $L_G = 25\,\text{nm}$. The counter doping of p-type and the sidewall spacer regions are added to manage the junction depletion effect. As can be seen from the drawing, the net doping switches from n-type to p-type outside of the intended body region ($0 \leq x \leq 25\,\text{nm}$). After all, the net doping in the body

region is p-type of concentration around $2.2 \times 10^{18}\,\mathrm{cm}^{-3}$.

The electron concentration in the 2D "slice" at various bias voltages (V_{GS} and V_{DS}) is shown in Figure 6.3. In addition, the electron concentration at the surface ($y = 0\,\mathrm{nm}$) and the middle of the body ($y = 5\,\mathrm{nm}$) is shown in Figure 6.4. We will discuss the effects on channel inversion from drain bias (V_{DS}) and gate bias (V_{GS}) in this order.



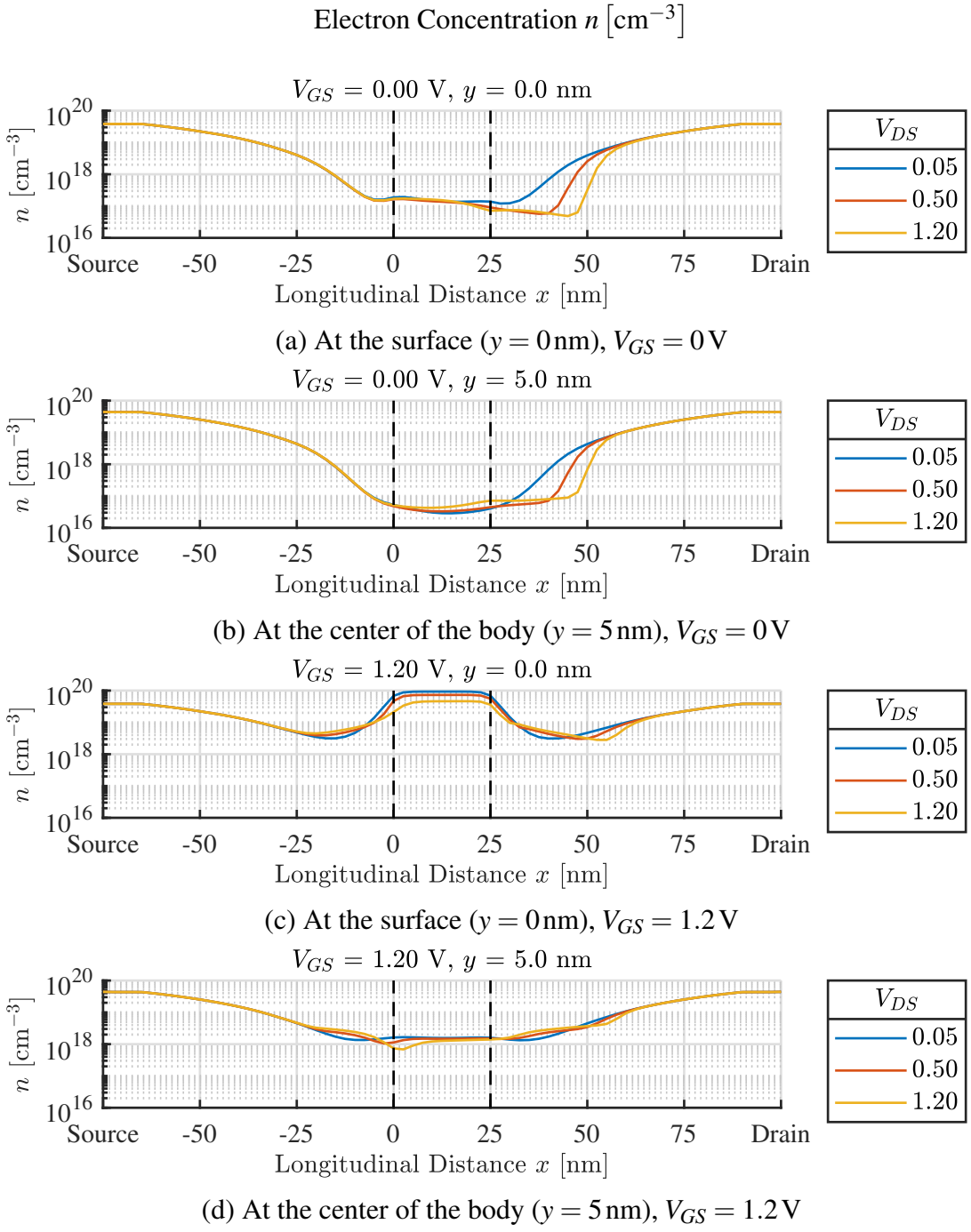
(a) $V_G = 0.0\,\mathrm{V}, V_D = 0.0\,\mathrm{V}$

(b) $V_G = 0.0\,\mathrm{V}, V_D = 0.8\,\mathrm{V}$

(c) $V_G = 0.8\,\mathrm{V}, V_D = 0.0\,\mathrm{V}$

(d) $V_G = 0.8\,\mathrm{V}, V_D = 0.8\,\mathrm{V}$

Figure 6.3: Top-down cross-sectional view of FinFET electron concentration under various bias conditions. Only the body $0 \le x \le L_G = 25\,\mathrm{nm}$ and $0 \le y \le W_{BODY} = 10\,\mathrm{nm}$ and a part of the source, drain and spacers are shown. Colors show the concentration between $1 \times 10^{16}\,\mathrm{cm}^{-3}$ and $1.4 \times 10^{20}\,\mathrm{cm}^{-3}$ in log scale in all four plots.



(a) At the surface ($y = 0$ nm), $V_{GS} = 0$ V

(b) At the center of the body ($y = 5$ nm), $V_{GS} = 0$ V

(c) At the surface ($y = 0$ nm), $V_{GS} = 1.2$ V

(d) At the center of the body ($y = 5$ nm), $V_{GS} = 1.2$ V

Figure 6.4: Electron concentration in the FinFET (the "basic" design) body/channel region at the surface ($y = 0$ nm) and the middle of the body ($y = 5$ nm), under various drain terminal bias conditions ($V_{DS} = 0.05$ V, 0.5 V, and 1.2 V; $V_{GS} = 0$ V and 1.2 V). The channel region is $0 \leq y \leq 25$ nm enclosed by the dashed lines.

Under zero gate bias ($V_{GS} = 0\,\mathrm{V}$), increasing V_{DS} causes less surface inversion ($y = 0\,\mathrm{nm}$) around the channel end on the drain side ($x = 25\,\mathrm{nm}$), from $n = 1.4 \times 10^{17}\,\mathrm{cm}^{-3}$ at $V_{DS} = 0.05\,\mathrm{V}$ to $n = 7.1 \times 10^{16}\,\mathrm{cm}^{-3}$ at $V_{DS} = 1.20\,\mathrm{V}$. In contrast, at the center of the body ($y = 5\,\mathrm{nm}$), the concentration increases from $n = 4.1 \times 10^{16}\,\mathrm{cm}^{-3}$ at $V_{DS} = 0.05\,\mathrm{V}$ to $n = 7.1 \times 10^{16}\,\mathrm{cm}^{-3}$ at $V_{DS} = 1.20\,\mathrm{V}$.

With gate voltage $V_{GS} = 1.20\,\mathrm{V}$, the concentration in the channel ($0\,\mathrm{nm} \le x \le 25\,\mathrm{nm}$) at the surface ($y = 0\,\mathrm{nm}$) varies from $n = 9.3 \times 10^{19}\,\mathrm{cm}^{-3}$ at $V_{DS} = 0.05\,\mathrm{V}$ to $n = 4.6 \times 10^{19}\,\mathrm{cm}^{-3}$ at $V_{DS} = 1.20\,\mathrm{V}$, showing the depletion effect. The concentration at the middle of the body ($y = 5\,\mathrm{nm}$) under the same gate bias is, contradicting to the zero-bias scenario, almost constant at the channel end near the drain (stays $\sim 1.5 \times 10^{18}\,\mathrm{cm}^{-3}$), but at the source end ($x = 0\,\mathrm{nm}$), it decreases from $n = 1.6 \times 10^{18}\,\mathrm{cm}^{-3}$ at $V_{DS} = 0.05\,\mathrm{V}$ to $n = 7.4 \times 10^{17}\,\mathrm{cm}^{-3}$ at $V_{DS} = 1.20\,\mathrm{V}$.

In summary, the effect of the drain bias voltage V_{DS} on the channel electron concentration is generally insignificant. For reference, some often-interested length quantities are listed in Table 6.3.

Table 6.3: Some Length Quantities of the Simulated FinFET

Quantity	Expression	Length	Note
Debye length	$L_D = \sqrt{\frac{\varepsilon_{Si} V_T}{qn}}$	4.1 nm	$n = 1 \times 10^{18}\,\mathrm{cm}^{-3}$ is used
Abrupt junction depletion region width	$x_p = \sqrt{\frac{2\varepsilon_{Si}}{q} \frac{N_D}{N_A} \frac{1}{N_D + N_A} \phi_0}$ (Equation 6.2)	25 nm	Mostly contained in the spacer L_{SP}
Natural length	$\lambda = \sqrt{\frac{1}{N_G} \frac{\varepsilon_{Si}}{\varepsilon_{OX}} W_{BODY} t_{OX}}$ (Equation 6.1)	6.7 nm	$N_G = 2$ is used for the 2D "slice"

From the body doping concentration, one may infer that the device has a positive threshold voltage, since from the textbook expression

$$V_{TH} = \frac{1}{C_{OX}} \sqrt{4q\varepsilon_{Si} N_A \left|\phi_p\right|} + 2\left|\phi_p\right| - (\phi_n - \phi_p) = 0.69\,\mathrm{V} \tag{6.3}$$

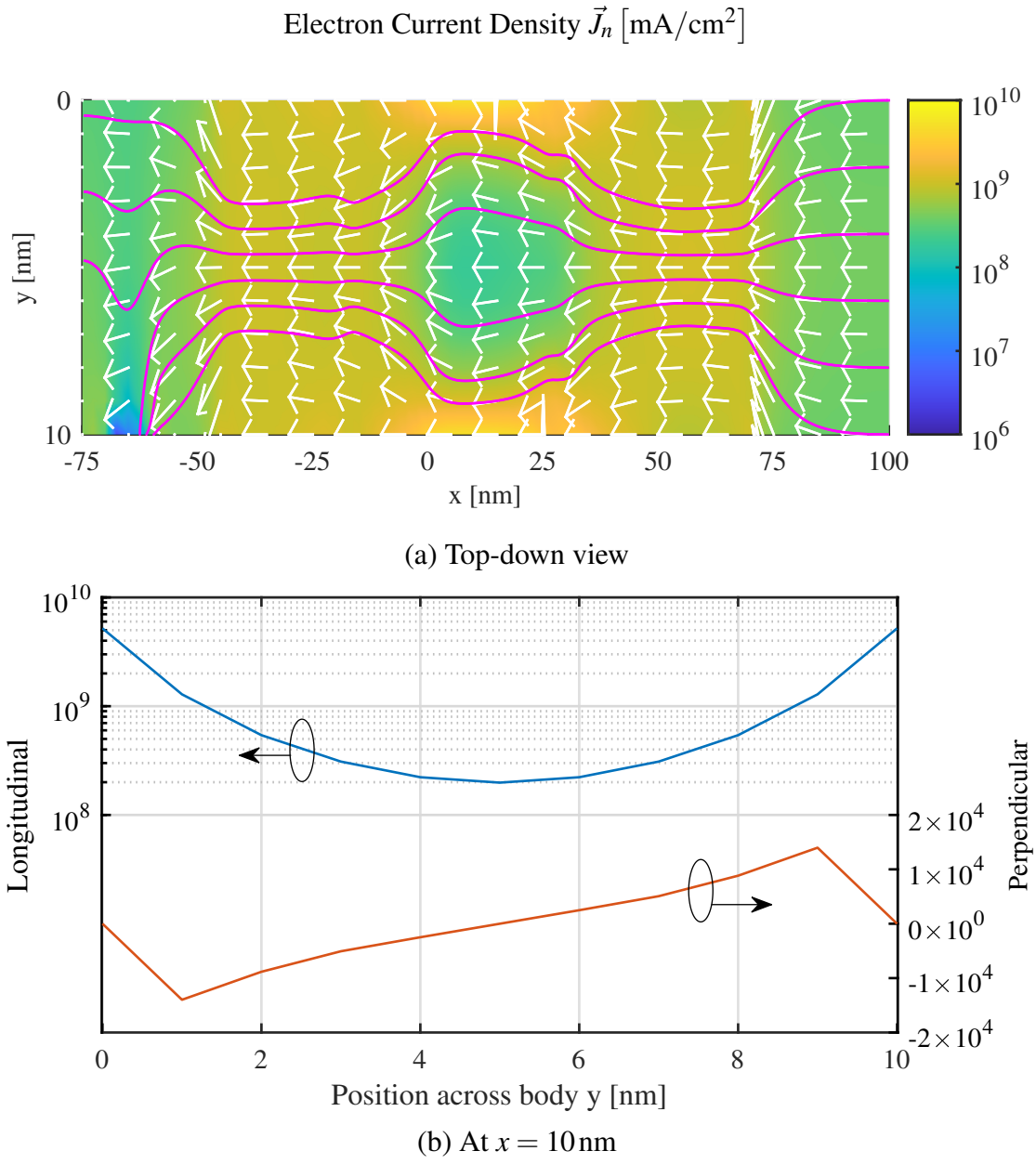
where $C_{OX} = \frac{\varepsilon_{OX}}{t_{OX}} = 1.15\,\mu\mathrm{F/cm}^2$ for the gate oxide, $\phi_p = -V_T \ln \frac{N_A}{n_i} = -0.50\,\mathrm{V}$ for the substrate, and $\phi_n = \frac{1}{2} \frac{\mathscr{E}_g}{q} = 0.55\,\mathrm{V}$ for the gate electrode. However, from the plots in Figure

6.3, an abundance of electrons (more than $1 \times 10^{16}\,\mathrm{cm}^{-3}$) is always present under non-negative V_G, although a significant increase can be seen as V_G increases from 0 V to 0.8 V.

It is believed that, because of the narrow body and the workfunction difference between the gate electrode and body (substrate), the entire body is depleted and inverted even under zero bias [111, 112]. But the electron concentration at zero bias is not enough to provide significant channel conduction, so the threshold voltage is still positive $V_{TH} > 0$.

On the other hand, since the whole body region $0 \leq y \leq W_{BODY} = 10\,\mathrm{nm}$ is capable of channel conduction, this is advantagous for the ON state operation in practical uses. With a proper device design and die layout, one may achieve higher current density than using planar devices in a same wafer area; effectively, R_{DSon} may be reduced.

In Figure 6.5a, the current density in the top-down view at forward bias is shown. In Figure 6.5b, the current density at $x = 10\,\mathrm{nm}$ indicates that the channel current at the center of the body, which is the smallest reading ($2.0 \times 10^8\,\mathrm{mA/cm^2}$) across the body, is about 3.9 % of that at the surface, which is the highest reading ($5.2 \times 10^9\,\mathrm{mA/cm^2}$).



(a) Top-down view

(b) At $x = 10\,\text{nm}$

Figure 6.5: Top-down cross-sectional view of FinFET electron current density under bias $V_G = 0.8\,\text{V}, V_D = 0.8\,\text{V}$.
In (a), the body region $0 \le x \le L_G = 25\,\text{nm}$ and $0 \le y \le W_{BODY} = 10\,\text{nm}$ and a part of the source, drain and spacers are shown. Colors show the magnitude $\left|\vec{J}_n\right|$, arrows show the vector directions at places where the tails are, and streamlines trace the flow starting from the drain electrode.
In (b), the longitudinal (along-channel) and perpendicular (towards the semiconductor-oxide interfaces) current densities at $x = 10\,\text{nm}$ are shown.

A separate simulation of a front-back cross-sectional view provides a second perspective to the device. The simulated device structure can be seen from the electric field-position plot shown in Figure 6.6. The new z direction is the height, perpendicular to the substrate wafer. The regions are: substrate $z \leq -10\,\text{nm}$; field oxide $-10\,\text{nm} \leq z \leq 0\,\text{nm}$, $-10\,\text{nm} \leq y \leq 0\,\text{nm}$ and $10\,\text{nm} \leq y \leq 20\,\text{nm}$; fin body $0\,\text{nm} \leq z \leq 34\,\text{nm}$, $0\,\text{nm} \leq y \leq 10\,\text{nm}$; gate oxide $0\,\text{nm} \leq z \leq 37\,\text{nm}$, $-3\,\text{nm} \leq y \leq 0\,\text{nm}$ and $0\,\text{nm} \leq y \leq 13\,\text{nm}$, and $34\,\text{nm} \leq z \leq 37\,\text{nm}$ and $0\,\text{nm} \leq y \leq 10\,\text{nm}$; the gate electrode (n++ poly-Si) is above the gate oxide. An ad-hoc doping concentration $N_A = 2.2 \times 10^{18}\,\text{cm}^{-3}$ is used for the substrate, according to the top-down simulation results when the drain and source implants are present. The substrate contact (boundary conditions) are set at the bottom ($z = -35\,\text{nm}$).

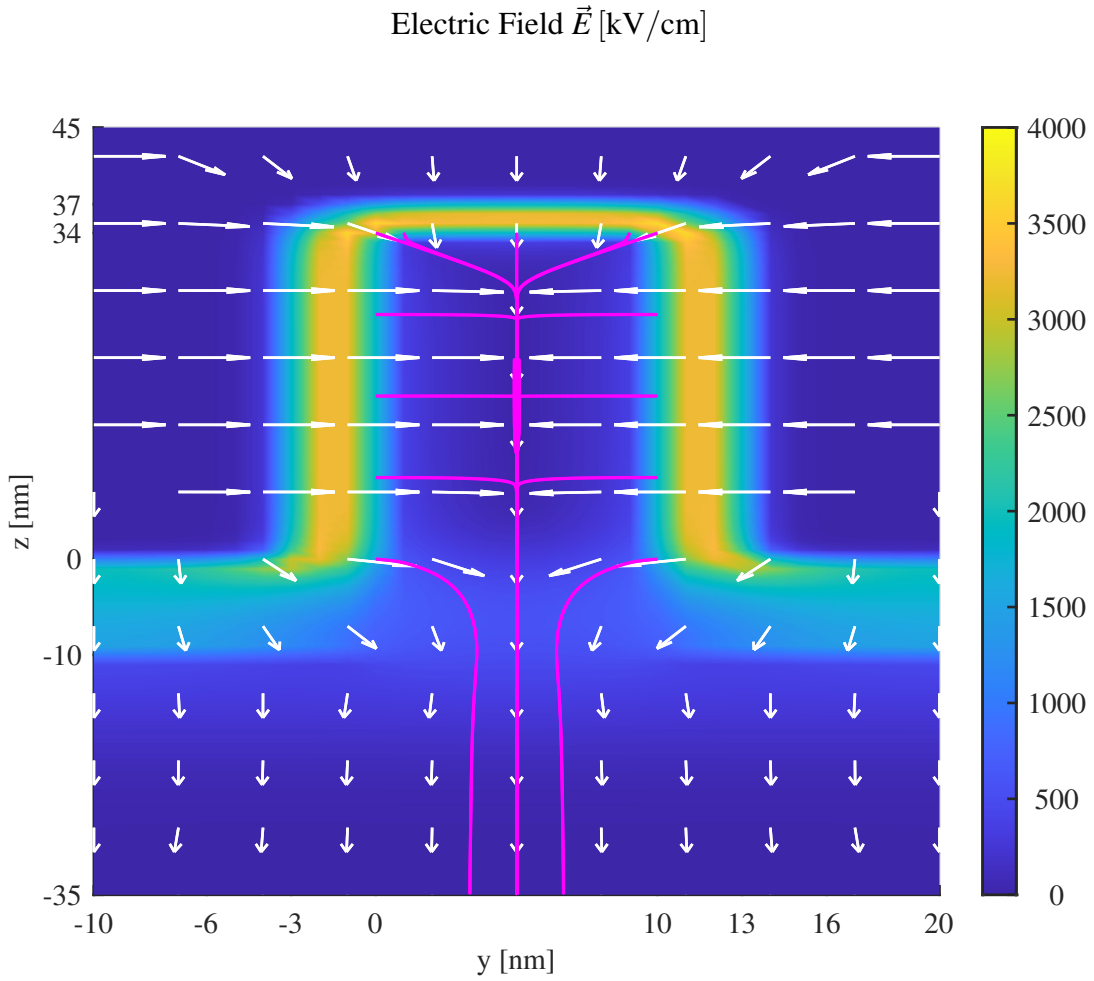


Figure 6.6: Front-Back Cross-Sectional View of the Electric Field $\vec{E}$ in the FinFET [kV/cm]. Bias $V_{GS} = 1.0\,\text{V}$. The entire simulation region is shown and described in the text. The colors indicate the local E field magnitude $\left|\vec{E}\right|$. The arrows show the directions of $\vec{E}$ where they start from. The streamlines trace the field starting from the semiconductor-oxide interface.

From the plot, we can observe that the field in the fin body generally points from the gate-oxide interface to the substrate and to the center of the body. Thus, the potential generally decreases in the same directions, and so does the electron concentration, which is sampled at various locations ($z = -5\,\text{nm}$ below the field oxide, 5, 20, and $34\,\text{nm}$ in the fin; $y = 0$, 1, 3, and $5\,\text{nm}$ across the fin) and plotted against sweeping V_{GS} in Figure 6.7. After closely examining the data, we can summarize that the electron concentration n under the

same V_{GS} is higher where: 1) closer to the gate-oxide interface (y decreases), and 2) further protruding away from the substrate (z increases). Besides, when V_{GS} increases, n increases.

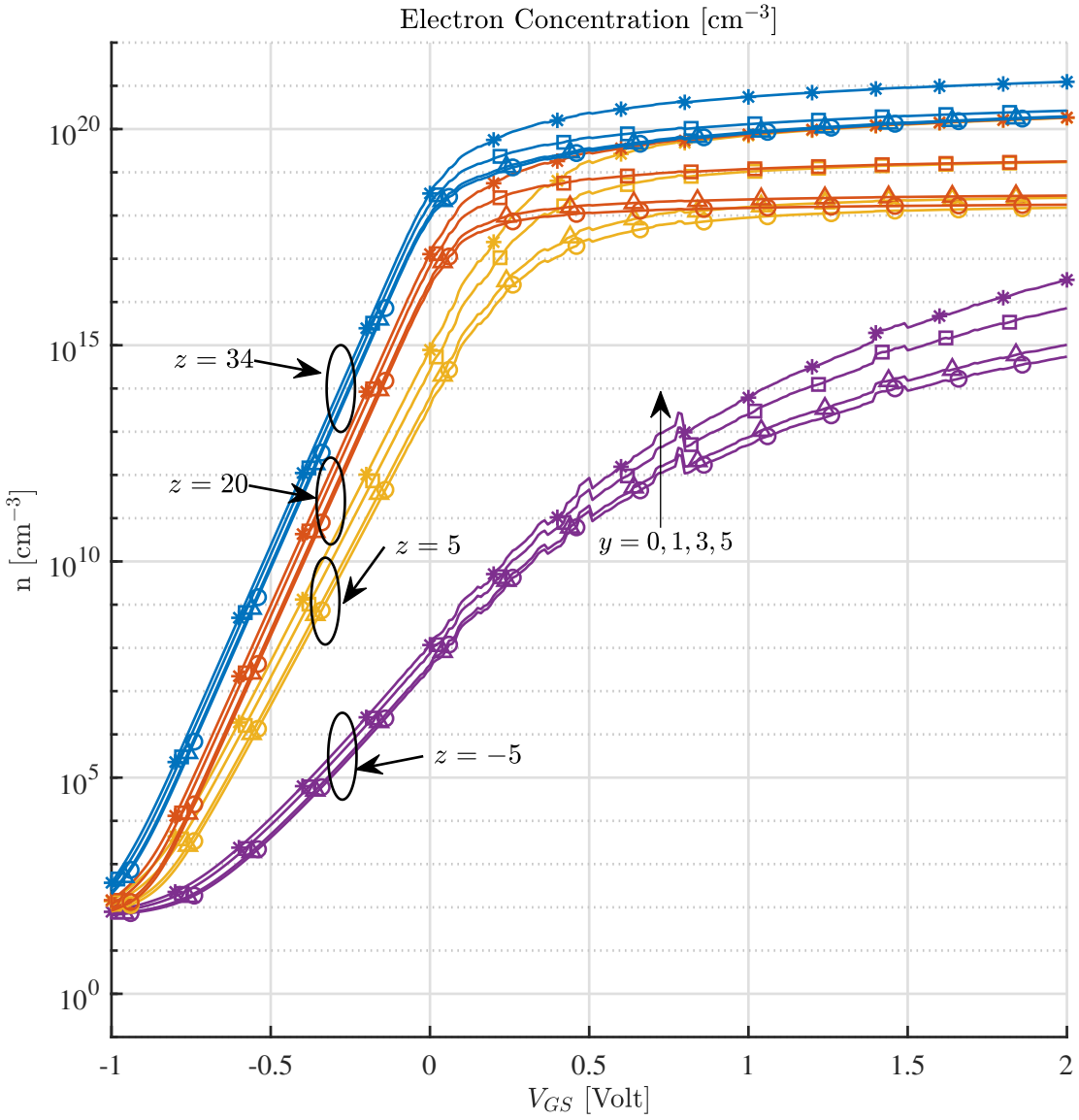
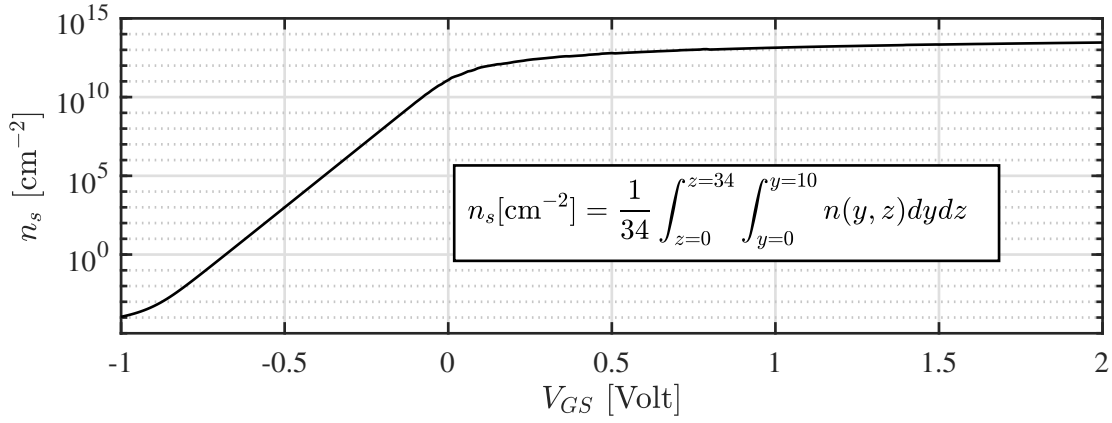
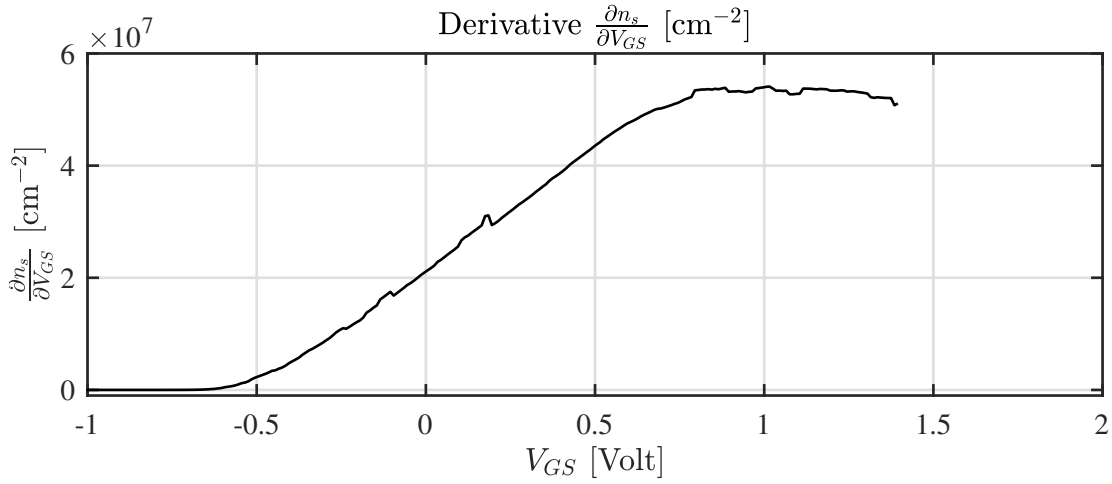


Figure 6.7: Electron concentration in the front-back simulation of the FinFET $n(y,z)\left[\text{cm}^{-3}\right]$ under changing bias voltage V_{GS}. The sample locations are a mesh grid with $z = -5$, 5 , 20 , and 34 nm in height; $y = 0$, 1 , 3 , and 5 nm in width.

The concentration-bias dependency can be simplified into an spatial average value, or area electron concentration n_s, by integrating $n(y,z)$ over the effective body region $0 \leq z \leq 34\,\text{nm}, 0 \leq y \leq 10\,\text{nm}$ and dividing the result by the fin height $T_{BODY} = 34\,\text{nm}$. The results (in cm^{-2}) versus V_{GS} are shown in Figure 6.8a.



(a) Area electron concentration n_s $\left[\mathrm{cm}^{-2}\right]$



(b) Derivative $\frac{\partial n_s}{\partial V_{GS}}$ $\left[\mathrm{cm}^{-2}\,\mathrm{V}^{-1}\right]$

Figure 6.8: (a) Area electron concentration n_s $\left[\mathrm{cm}^{-2}\right]$ in the fin body under sweeping V_{GS}, by integrating the volume concentration n $\left[\mathrm{cm}^{-3}\right]$ over the body cross section (height $0 \leq z \leq 34$ nm, width $0 \leq y \leq 10$ nm) and dividing it by the fin height 34 nm. (b) its derivative versus bias voltage $\frac{\partial n_s}{\partial V_{GS}}$ $\left[\mathrm{cm}^{-2}\,\mathrm{V}^{-1}\right]$.

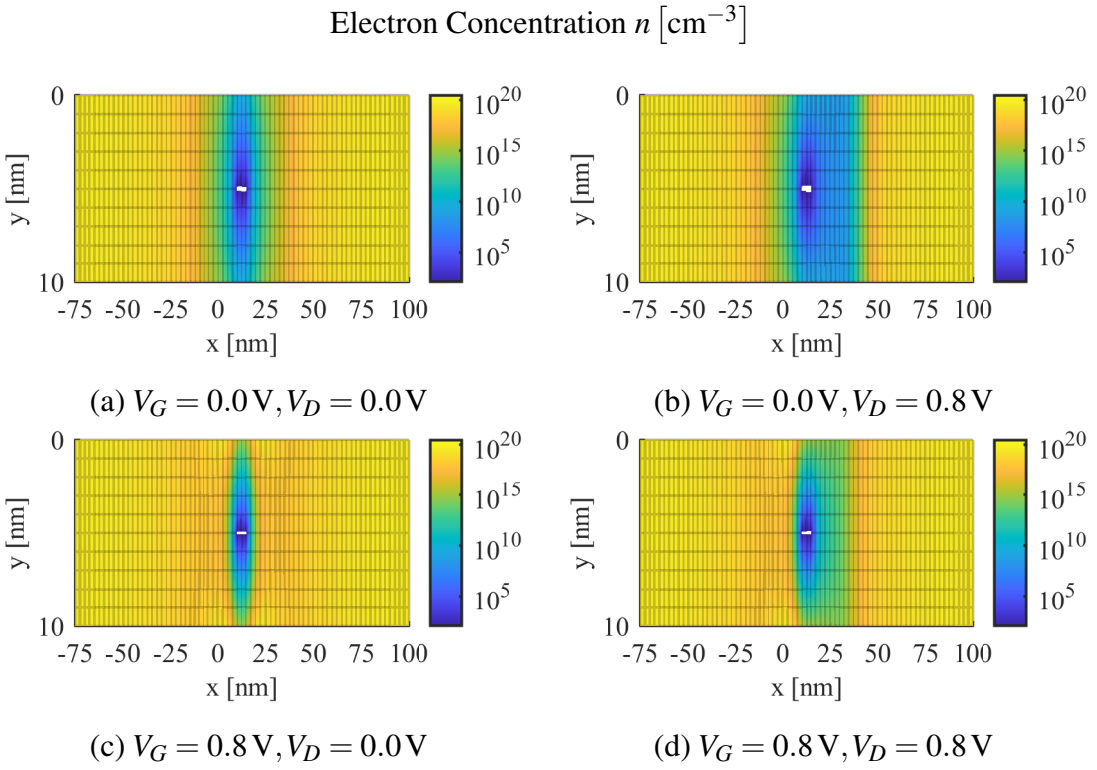
Since the channel current in the linear region roughly linearly depends on the channel carrier concentration[20], we can apply the ELR method and extract an estimated threshold voltage $V_{TH} = 0.13\,\mathrm{V}$. The derivative used is $\frac{\partial n_s}{\partial V_{GS}}$, as shown in Figure 6.8b, instead of the normally used transconductance $G_m = \frac{\partial I_D}{\partial V_{GS}}$.

However, according to the results from using the compact model (BSIM-CMG), as will be discussed in the next section, the results shown here may not be realistic, i.e., the

[20]Since in the linear region (low drain-source voltage), $I_D \approx \mu_n Q_{CH} \frac{V_{DS}}{L} \propto Q_{CH} = q n_s$.

body may not be under inversion at low bias. An example **incorrect** way to artificially make V_{TH} higher is by placing the unused body electrode (defining boundary conditions) in the middle of the body, in the top-down view. It sets up an additional set of boundary conditions and lowers the body potential generally. By doing so, not only is the potential forced to a given value at zeroth order (Dirichlet boundary conditions) but also the body electrode can "steal" the channel current from the source, as can be observed from I-V characteristics. A 1 kΩ resistor is inserted between the body terminal and circuit ground, representing the resistance of the substrate between the body region (fin) and the "body" terminal contact and preventing the body current mentioned above due to the improper boundary conditions.

The electron concentrations at various V_G with the additional body electrode in the middle of the body are shown in Figure 6.9, tagged "alternative design". The body electrode boundary condition is set at $10\,\text{nm} \leq x \leq 15\,\text{nm}$ and $y = 5\,\text{nm}$ (the blank gaps in the plots), and the electrode's workfunction is set to an ad-hoc value of 5.12 eV, which refers to the equilibrium Fermi level, relative to the vacuum energy, when $p = 5 \times 10^{18}\,\text{cm}^{-3}$. Partial inversion can still be observed in the channel under zero bias, but much delayed.



(a) $V_G = 0.0\,\mathrm{V}, V_D = 0.0\,\mathrm{V}$

(b) $V_G = 0.0\,\mathrm{V}, V_D = 0.8\,\mathrm{V}$

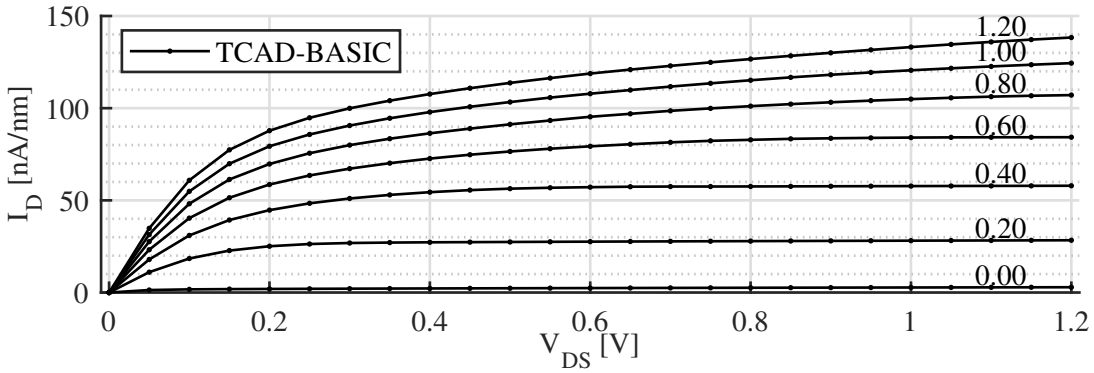
(c) $V_G = 0.8\,\mathrm{V}, V_D = 0.0\,\mathrm{V}$

(d) $V_G = 0.8\,\mathrm{V}, V_D = 0.8\,\mathrm{V}$

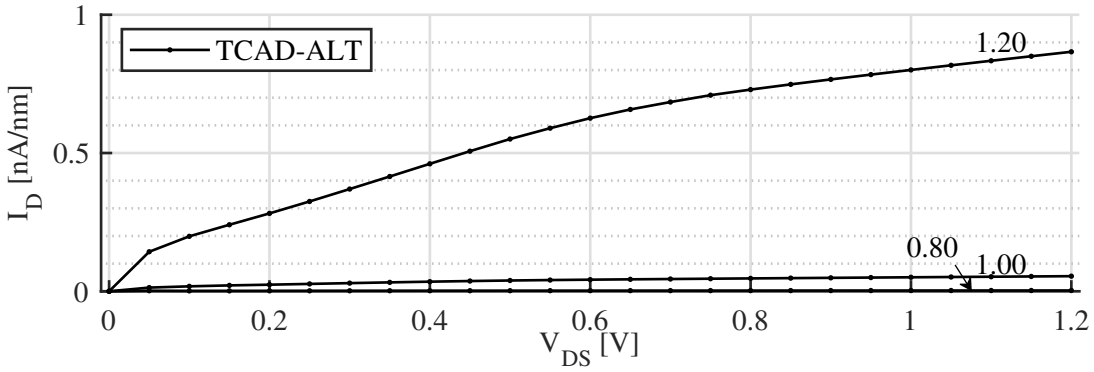
Figure 6.9: Top-down cross-sectional view of FinFET (alternative design, additional electrode boundary conditions in the middle of the body) electron concentration under various bias conditions. Only the body $0 \leq x \leq L_G = 25\,\mathrm{nm}$ and $0 \leq y \leq W_{BODY} = 10\,\mathrm{nm}$ and a part of the source, drain and spacers are shown. Colors show the concentration between $1 \times 10^2\,\mathrm{cm}^{-3}$ and $1.4 \times 10^{20}\,\mathrm{cm}^{-3}$ in log scale in all four plots.

The I-V characteristics of the simulated FinFET device (2D cross-section) are shown in Figure 6.10 (I_D-V_{DS}) and Figure 6.11 (I_D- and G_m-V_{GS}). In Figures 6.10a, 6.10b, 6.11a, 6.11b, 6.11c, and 6.11d, the mobility model parameters extracted from the planar MOSFET model in Section 3.2 are applied, assuming the mobility-related material properties do not deviate by too much between the FinFET structure and the conventional planar MOSFET structure. These custom parameters are then removed, and thus the default mobility model defined in Cider is used. The results can be seen in Figures 6.10c, 6.11e and 6.11f. In all simulations shown here, local carrier generation due to impact ionization is not calculated.

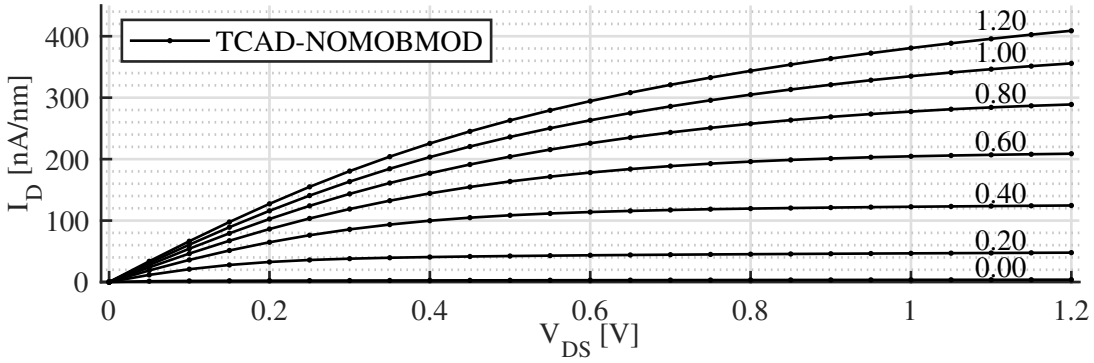
Drain Current per Fin Height I_D-V_{DS} /T_{BODY} [nA/nm]



(a) "Basic" design

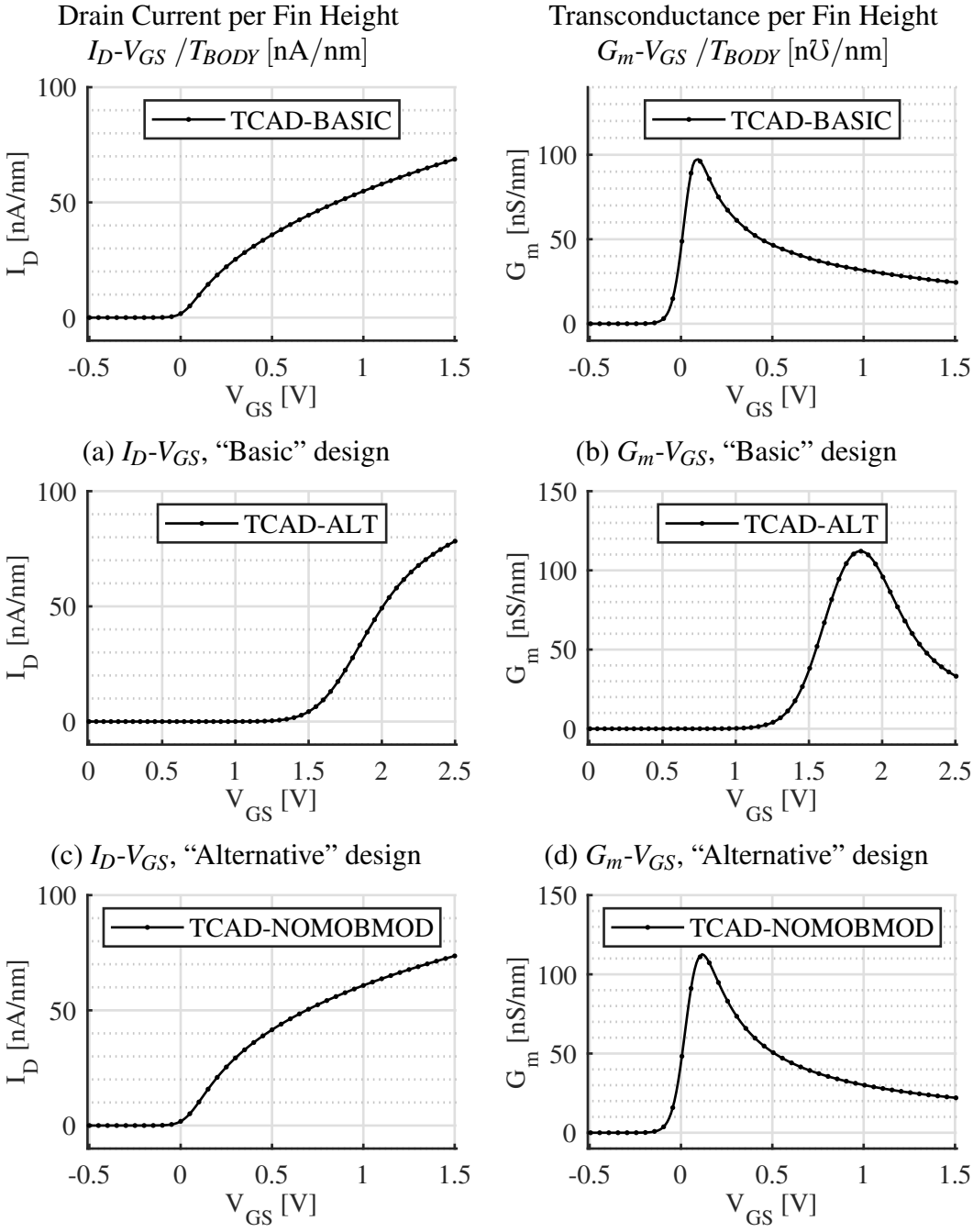


(b) "Alternative" design



(c) "Basic" design with built-in mobility model

Figure 6.10: FinFET terminal I_D-V_{DS} per fin height (T_{BODY}), from the top-down cross-sectional device simulation under various V_{GS} voltages as annotated next to the lines. The current readings are converted to per-unit-fin-height values. Three designs are included: (a) The "basic" design introduced at the beginning of Section 6.2, (b) The "alternative" design with additional boundary conditions, and (c) The "basic" geometric and doping configuration, but with Cider's default mobility model. The mobility model used in the first two cases are the one extracted from Section 3.2.



(a) I_D-V_{GS}, "Basic" design

(b) G_m-V_{GS}, "Basic" design

(c) I_D-V_{GS}, "Alternative" design

(d) G_m-V_{GS}, "Alternative" design

(e) I_D-V_{GS}, "Basic" design with built-in mobility model

(f) G_m-V_{GS}, "Basic" design with built-in mobility model

Figure 6.11: FinFET terminal I_D-V_{GS} per fin height (T_{BODY}), from the top-down cross-sectional device simulation with $V_{DS} = 0.1\,\mathrm{V}$. The current readings are converted to per-unit-fin-height values. Three designs are included, same as those in Figure 6.10.

From the I_D-V_{GS} data, the threshold voltages V_{TH} of the three device designs are extracted using the ELR method [78], with the calculated $G_m = \frac{\partial I_D}{\partial V_{GS}}$ plotted aside in Figure 6.11. They are 0.10 V ("basic"), 0.30 V ("alternative"), and 0.11 V ("basic" with built-in mobility model) for the three designs, respectively. For comparison, the front-back cross-sectional simulation yields an estimated $V_{TH} = 0.13$ V.

With the above results in mind, the following summaries can be drawn by comparing the three design cases. Apparently, after applying the extracted mobility model, the drain current under strong inversion is decreased to $\sim 30\,\%$ of the result using the default model built into Cider, while the validity of projecting the surface mobility model extracted from a planar, long-channel device to the FinFET under investigation remains questionable. But the extracted threshold voltage and the subthreshold drain current do not seem to be affected as much (the maximum transconductance has decreased by $\sim 14\,\%$). Meanwhile, introducing the additional boundary conditions significantly lowers the body potential, lowers the drain current under strong inversion, and increases the threshold voltage as expected. However, when the *real* body electrode is present in the front-back cross-sectional simulation, V_{TH} is around 0.1 V too, in contrast to the "alternative" design. Therefore, we cannot conclude that it is a proper improvement by applying an additional set of boundary conditions in the center of the body.

6.3 Mesoscopic Simulation (1D) of FinFETs — a Quantum-Corrected Solution

In this section, the quantum-mechanical effects in the FinFET's body (Si) are investigated. In a small device like the FinFET discussed in this work, quantum effects can become significant. As a result, the solution (e.g., carrier concentration, band structure, E field) may differ from the classical version.

The FinFET cross section has an apparent resemblance to the classic case of a particle in a box or finite-barrier potential well problem set-up, leading to sinusoidal wavefunctions with integer numbers of half-waves in the body as the solution. Additionally, the band bending effect near the semiconductor-oxide interface gives rise to triangular potential wells at the interfaces. The wavefunction solutions to the ideal infinite triangular well problem are Airy functions [113]; the length scale of these triangular wells and wavefunctions can be estimated by the Debye length and the natural length, which are both about half of the studied FinFET's body width, according to Table 6.3.

The actual solution to the FinFET body should be a combination of the above two, suggesting that the electron concentration, determined by the wavefunction magnitudes, reaches its peak value away from the interface. The channel carrier concentration under inversionat the interface (i.e., electrons in an n-type device) is significantly reduced. In other words, the channel electrons are "pushed away" from the interfaces into the body due to quantum-mechanical effects. This contrasts with the classical solution to the device-level semiconductor equations, or Poisson's equation (Equation 2.25a) under quasi-equilibrium conditions, where the channel carrier concentration typically peaks at the semiconductor-oxide interface.

The rest of this section consists of four parts — **1)** An outline of our simulation methodology and a concise introduction to our implementation of the finite difference method; **2)** The detailed derivation of the system of equations in one-dimensional space,

where we define and describe most concepts and quantities omitted in the outline; **3)** Simulation results of the FinFET and discussion; and **4)** Extracting the gate oxide field for oxide breakdown modeling in the next step (Section 6.4).

6.3.1 Methodology Outline

We set up a one-dimensional finite potential well structure to represent the FinFET structure. As position y increases (going from left to right), five regions are present: n-type Si gate terminal, SiO_2 gate dielectric layer (potential barrier), p-type Si body, SiO_2 gate dielectric layer (potential barrier), and n-type Si gate terminal. The system is presumed to be quasi-equilibrium, and the input is the gate terminal voltage V_{GS}.

Our method to solve the system of equations for potential $\phi(y)$, electron concentration $n(y)$, as well as allowed energy states $\mathscr{E}_m(y)$ and their associated wavefunctions $\Psi_m(y)$ is illustrated with the flow chart in Figure 6.12.

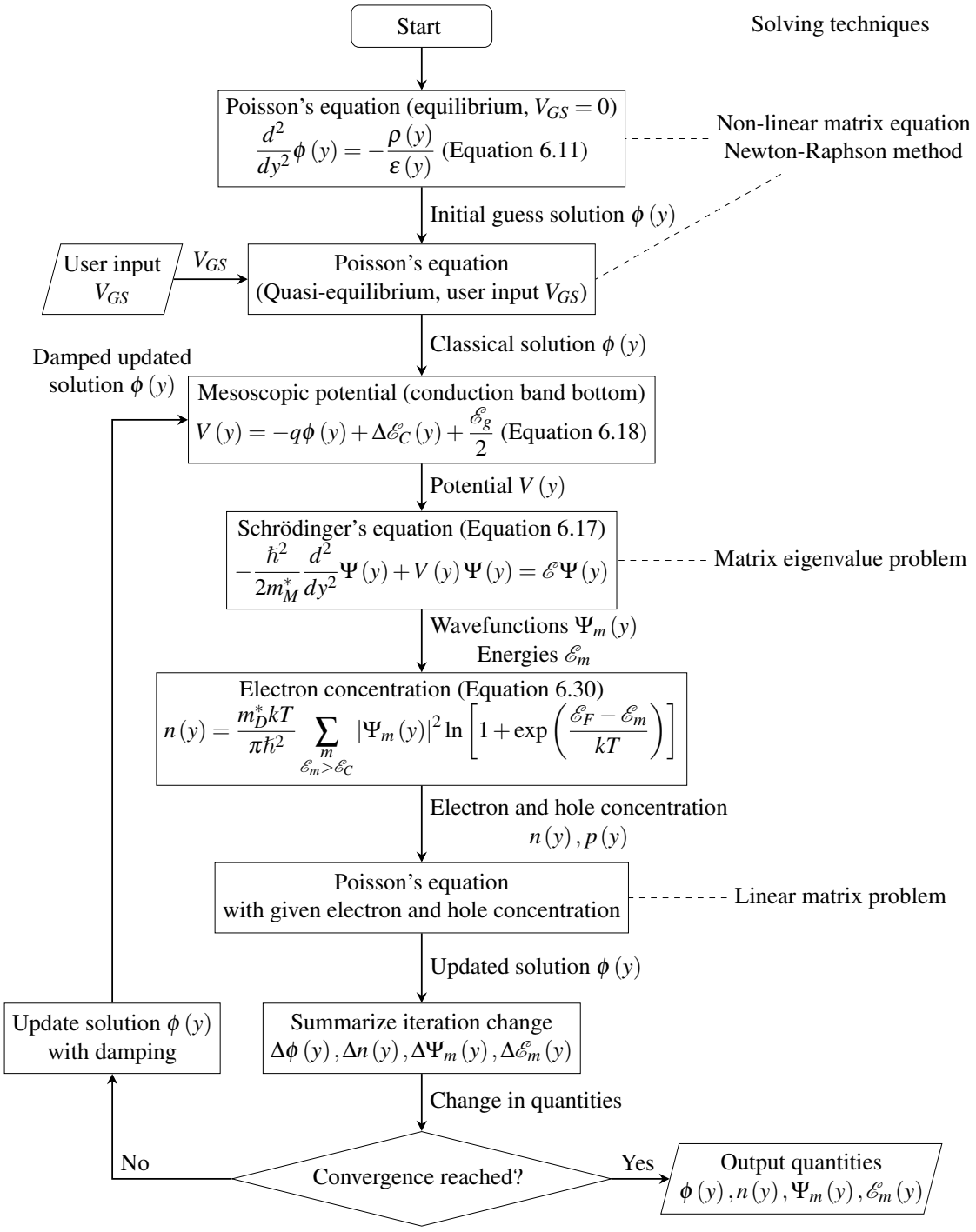


Figure 6.12: A flowchart showing our approach of applying quantum correction to the FinFET device simulation.

All equations and quantities are solved numerically using finite difference. In our 1D representation, for example, the space y becomes a discretized variable after applying finite difference as:

$$y_i, \qquad i = 0, 1, 2, \ldots, N \tag{6.4}$$

Meanwhile, a continuous scalar function $f(y)$ is represented as:

$$f_i = f(y_i), \qquad i = 0, 1, 2, \ldots, N \tag{6.5}$$

while its derivatives are the finite differences given as:

$$\left.\frac{df}{dy}\right|_{y=y_i} = \frac{f_{i+1} - f_i}{\Delta y} \qquad \text{(Forward difference)} \tag{6.6}$$

$$\left.\frac{df}{dy}\right|_{y=y_i} = \frac{f_i - f_{i-1}}{\Delta y} \qquad \text{(Backward difference)} \tag{6.7}$$

$$\left.\frac{d^2 f}{dy^2}\right|_{y=y_i} = \frac{f_{i+1} - 2f_i + f_{i-1}}{(\Delta y)^2} \qquad \text{(Central difference)} \tag{6.8}$$

where $\Delta y = y_{i+1} - y_i$ is the constant spatial interval in our uniformly distributed spatial positions ("mesh points").

First, we solve the Poisson's equation given below (and Equation 6.11) for the electrostatic potential $\phi(y)$ under applied voltage V_{GS}:

$$\frac{d^2}{dy^2}\phi(y) = -\frac{\rho(y)}{\varepsilon(y)} \tag{6.9}$$

where the position-dependent charge density $\rho(y)$ is unknown and depends on the potential, and $\varepsilon(y)$ is the material-dependent dielectric constant.

In this step, the Poisson's equation is solved by using the Newton-Raphson method (or "Newton's method"), which is a gradient descent method. The Newton-Raphson method iteratively evaluates the local gradient, and computes a "step" towards an improved solution $\Delta\phi_i = \Delta\phi(y_i)$ with respect to all mesh points y_i, until convergence is reached, i.e., its magnitude[21] $|\Delta\phi(y)|$ ("step size") becomes smaller than a pre-defined threshold (0.01% of the magnitude of $\phi(y)$ in our case). A separate solution for the identical system but

[21] Considering the discretized space $y_i\,(i = 0, 1, 2, \ldots, N)$, the function $\Delta\phi(y_i) = \Delta\phi_i$ can be treated as a vector in the finite-difference space (i.e., not the physical space), and its magnitude is the vector norm or the geometric average of the values at all mesh points: $|\Delta\phi| = \left[\sum_i (\Delta\phi_i)^2\right]^{1/2}$.

with zero input voltage ($V_{GS} = 0$) is calculated beforehand to be used as the initial guess when $V_{GS} \neq 0$. Also, the "improvement" variable $\Delta\phi(y)$ may be scaled down in each iteration step in order to avoid failures to converge (because the system is too far away from equilibrium), at the price of having more iterations.

Once we have acquired the classical solution from the quasi-equilibrium Poisson's equation, we apply quantum correction by adding three equations to the system.

First, the mesoscopic potential $V(y)$, which describes the potential well formed by the oxide-semiconductor-oxide structure, is calculated from the classical electrostatic potential $\phi(y)$ and material properties, as defined later in Equation 6.18.

With this potential, we can set up the time-independent Schrödinger's equation given as below (and in Equation 6.17):

$$-\frac{\hbar^2}{2m_M^*}\frac{d^2}{dy^2}\Psi(y) + V(y)\Psi(y) = \mathscr{E}\Psi(y) \tag{6.10}$$

Bound states are expected in the finite well; therefore, we apply the Dirichlet boundary conditions (zero wavefunction at both ends, both of which are deep into the n-type Si gate terminals) and solve the eigenvalue problem associated with the Schrödinger's equation in its finite-difference matrix form.

With the eigenvalue solutions (i.e., allowed energy states $\mathscr{E}_m$ and their associated wavefunctions $\Psi_m(y)\,(m = 1, 2, \ldots)$), we calculate the quantum-corrected electron concentration $n(y)$ using Equation 6.30.

Finally, with the "improved" or quantum-mechanical solution of charge density $\rho(y)$, we solve the Poisson's equation again for the electrostatic potential $\phi(y)$. Note that in this step, the only unknown is $\phi(y)$, and the Poisson's equation in finite-difference form is a linear algebraic system of equations and directly solvable by matrix inversion or other equivalent techniques.

With the updated solutions with quantum correction: $\phi(y)$, $\Psi_m(y)$, $\mathscr{E}_m$, and $n(y)$, we check all quantities for convergence; the process will be described in details later. If convergence is not yet reached, we go back to the step evaluating the mesoscopic potential

$V(y)$, replace the previously used electrostatic potential solution $\phi(y)$ with the updated one, and repeat the evaluation.

Once convergence is accomplished, we have the quantum-corrected solution to all quantities: the electrostatic potential $\phi(y)$, the mesoscopic potential (i.e., conduction band bottom) $V(y)$, the electron conecntration $n(y)$, as well as all bound states (wavefunctions $\Psi_m(y)$ and their associated energy eigenvalues $\mathscr{E}_m$).

The calculated electric field with quantum correction will be used for oxide breakdown analysis shortly. For example, we can apply the gate voltage $V_{GS} = V_{BD}$ according to the RBD condition from our experiments, and extract E_{BD} from the simulation result.

6.3.2 System of Equations

The problem is set up as follows. The semiconductor (Si) system is the FinFET structure in a 1D cross section view through the fin.[22] In the one-dimension space, the "cross section" is a line cutting through the gate-body-gate structure. In the top-down view in Figure 6.2, the cross section is taken at $x = 12.5\,\text{nm}$ in the middle of the body. Included structures are the p-type silicon substrate ($N_A = 2.2 \times 10^{18}\,\text{cm}^{-3}$), gate oxide layers ($SiO_2$) and n-type gate electrode ($N_D = 1 \times 10^{20}\,\text{cm}^{-3}$). We consider the quasi-equilibrium scenario, which means the system may be subject to gate bias voltages, but there is no conductive current, generation, or recombination. The classical system of equations in Equations 2.25a-2.25e reduces to just Poisson's equation

$$\frac{d^2}{dy^2}\phi(y) = -\frac{q}{\varepsilon(y)}\left[p(y) - n(y) + N_D(y) - N_A(y)\right] \tag{6.11}$$

The electrostatic potential $\phi(y)$ is defined as

$$\phi(y) = \frac{\mathscr{E}_i(y) - \mathscr{E}_F(y)}{q} \tag{6.12}$$

In other words, it is the difference between the intrinsic Fermi level $\mathscr{E}_i$ and the actual Fermi level $\mathscr{E}_F$. $\mathscr{E}_i$ is approximately in the middle of the bandgap; band bending effect is included. $\mathscr{E}_F$ is set to zero for the substrate, and $\mathscr{E}_F = -qV_{GS}$ for the gate electrodes under forward bias V_{GS}. For the rest of the derivation, we omit the spatial dependency for $\mathscr{E}_F$ and imply $\mathscr{E}_F = 0$ in the body. Alternatively,

$$\phi(y) = \frac{\mathscr{E}_C(y) - \frac{1}{2}\mathscr{E}_g - \mathscr{E}_F}{q} \tag{6.13}$$

where $\mathscr{E}_C$ is the conduction band bottom.

The electron concentration n will be discussed in more details shortly. Meanwhile, since there are many more electrons than holes in the fin body under zero and forward bias

[22] The quantization effect in the perpendicular direction is observable but neglected in our case, because in the Π-shaped FinFET, the oxide layer on top of the fin is much thicker than those on both sides. Therefore, for our purpose of simulating the oxide BD condition, the perpendicular direction is not as important.

in the studied case, the hole concentration p is simplified to

$$p(y) = \frac{n_i^2}{n(y)} \tag{6.14}$$

by the law of mass action.

The material permittivity $\varepsilon(y)$ is a constant scalar for each material (Si and SiO_2 in our case). Since different materials are placed at various positions in our configuration, it is written as a position-dependent function in Equation 6.11 to emphasize the material dependency. However, it should not be confused with a variable permittivity value in one material. Therefore, the second-order derivative in Poisson's equation is a valid expression except at semiconductor-oxide interfaces, where $\varepsilon(y)$ abruptly changes going from one material to another; in the latter case, the boundary condition for perpendicular E-field components is applied instead, which is derived by integrating the Poisson's equation (representing Gauss's law[23]) once, and the result is shown later in Table 6.4.

We replace the classical formula for electron concentration n using the continuous density of state (DOS) function with the discretized DOS as follows [114]. The total volume electron concentration is the integral over the conduction band of the DOS function, multiplied by the occupancy probability, i.e.,

$$n(y) = \int_{\mathscr{E}_C}^{+\infty} d\mathscr{E}\, f(\mathscr{E})\, g(\mathscr{E}, y) \tag{6.15}$$

The state occupancy function $f(\mathscr{E})$ under quasi-equilibrium follows the Fermi-Dirac distribution

$$\begin{aligned} f(\mathscr{E}) &= f_{FD}(\mathscr{E}_y + \mathscr{E}_{xz}, \mathscr{E}_F) \\ &= \frac{1}{1 + \exp\left(\frac{\mathscr{E}_y + \mathscr{E}_{xz} - \mathscr{E}_F}{kT}\right)} \end{aligned} \tag{6.16}$$

where the total electron energy $\mathscr{E}$ is split into two terms $\mathscr{E}_y$ and $\mathscr{E}_{xz}$, because the system is supposed to have discrete and bound states in the y dimension, where the potential wells are present, while quasi-continuous or "continuous" states and energy levels are assumed in the other two dimensions where structural scales are relatively larger. The sigmoid function

[23] Gauss's law in 1D can be written as: $-\frac{d}{dy}\left[\varepsilon(y)\frac{d\phi(y)}{dy}\right] = \rho(y)$

f_{FD} only depends on the total energy $\mathscr{E} = \mathscr{E}_y + \mathscr{E}_{xz}$ ($\mathscr{E}_F = 0$ in the body). $k = 1.381\cdots\times 10^{-23}\,\mathrm{J\,K^{-1}}$ is the Boltzmann constant, and $T = 303\,\mathrm{K}$ is the lattice temperature.

In the y direction, the potential well is created by the oxide-substrate-oxide structure's conduction band bottom. The Schrödinger's equation for the 1D system and the mesoscopic potential energy V are given by

$$-\frac{\hbar^2}{2m_M^*}\frac{d^2}{dy^2}\Psi(y) + V(y)\Psi(y) = \mathscr{E}\Psi(y) \tag{6.17}$$

$$V(y) = -q\phi(y) + \Delta\mathscr{E}_C(y) + \frac{\mathscr{E}_g}{2} \tag{6.18}$$

In this potential well, discretized, bound states are allowed, with energies lower than the barrier, along with continuous, free states with higher energies. The allowed bound state energy eigenvalues $\mathscr{E}_m$ $(m = 1,2,3,\dots)$ and their associated wavefunctions $\Psi_m(y)$ are found by solving the Schrödinger's equation (Equation 6.17) in 1D as an eigenvalue problem. m_M^* is the mobility effective electron mass; $m_M^* = 0.26m_e$ is used for Si and $0.86m_e$ for SiO_2 (the vacuum electron mass $m_e = 9.109\cdots\times 10^{-31}\,\mathrm{kg}$). The mesoscopic potential $V(y)$ in Equation 6.18 is the conduction band bottom. The conduction band offset $\Delta\mathscr{E}_C$ is the difference between the affinities of SiO_2 and Si, given as

$$\Delta\mathscr{E}_C(y) = \begin{cases} 3.1\,\mathrm{eV} & -6\,\mathrm{nm} < x < -3\,\mathrm{nm} \text{ and } 13\,\mathrm{nm} < x < 16\,\mathrm{nm}\ (\text{in } SiO_2) \\ 0\,\mathrm{eV} & \text{elsewhere (in Si)} \end{cases} \tag{6.19}$$

Thus, the barrier height is 3.1 eV.

The DOS function $g(\mathscr{E}, y)$ can be split into two terms as well, i.e.,

$$g(\mathscr{E}, y) = g_y(\mathscr{E}_y, y)\ g_{xz}(\mathscr{E}_{xz}) \tag{6.20}$$

where g_y describes the discrete bound states confined by the potential well, and g_{xz} is for the continuous states in the other two directions[24]. For clarification, the DOS discussed below belongs to the Si body. $\mathscr{E}_y$ and $\mathscr{E}_{xz}$ are relative to the Si conduction band bottom or $V(y)$ in Equation 6.18. However, as will be seen later, the finite potential barriers allow non-zero electron wavefunctions in the gate dielectric (SiO_2) regions. Thus, the electrons

[24]Sometimes they are called "2D DOS" and "3D DOS", respectively. These terminologies are omitted here to avoid any confusion with the dimensions used in the simulation.

associated to the Si-region wavefunctions and energies can appear in the oxide. However, they should not be confused with the electrons in the SiO_2 conduction band (with energies higher than the 3.1 eV barrier and having their own DOS), which are extremely rare and neglected in the calculation.

For the discretized states at position y, the DOS is zero except for allowed energy levels $\mathscr{E}_m (m = 1,2,3,\dots)$, when it is determined by the quantum probability, or the wavefunction magnitude squared $|\Psi_m(y)|^2$. Therefore,

$$g_y(\mathscr{E}_y, y) = \sum_{\substack{m \\ \mathscr{E}_m > \mathscr{E}_C}} |\Psi_m(y)|^2 \, \delta(\mathscr{E}_y - \mathscr{E}_m) \tag{6.21}$$

which has unit $\mathrm{cm}^{-1}\,\mathrm{eV}^{-1}$.

The "continuous" DOS in x and z directions can be found similar to the textbook approach for the classical DOS expression. The two-dimensional area per k-state in a fictitious square crystal of side length L is $\left(\frac{\pi}{L}\right)^2$ under translational symmetric boundary conditions. The size of a quarter ring between wavenumbers k and $k+dk$ is $\frac{1}{4} \times 2\pi k dk \, (k \geq 0)$. Considering the spin degeneracy of two, the number of states in $(k, k+dk)$ per area L^2 is

$$N(k)\,dk = \frac{\frac{1}{4} \times 2\pi k dk}{\left(\frac{\pi}{L}\right)^2} \times 2 \times \frac{1}{L^2} = \frac{k dk}{\pi} \tag{6.22}$$

Rewrite the wavenumber in terms of energy, and we have

$$\begin{aligned} k &= \sqrt{\frac{2m_D^* \mathscr{E}}{\hbar^2}} \\ \frac{k dk}{\pi} &= \frac{1}{\pi} k(\mathscr{E}) \frac{dk}{d\mathscr{E}} d\mathscr{E} \\ &= \frac{1}{\pi} \sqrt{\frac{2m_D^* \mathscr{E}}{\hbar^2}} \sqrt{\frac{m_D^*}{\hbar^2}} \frac{1}{2\sqrt{\mathscr{E}}} d\mathscr{E} \\ &= \frac{m_D^*}{\pi \hbar^2} d\mathscr{E} \\ &= g_{xz}(\mathscr{E})\, d\mathscr{E} \end{aligned} \tag{6.23}$$

where m_D^* is the DOS effective electron mass in the x and z directions. In this study, $m_D^* = 1.08 m_e$ is used. Therefore the DOS function in x-z direction is

$$g_{xz} = \frac{m_D^*}{\pi \hbar^2} \tag{6.24}$$

which turns out to be independent of the energy $\mathscr{E}_{xz}$ in these two dimiensions. g_{xz} has unit $\mathrm{cm}^{-2}\,\mathrm{eV}^{-1}$.

Next, we put g_y, g_{xz}, and f back into Equation 6.15 which yields the volume electron concentration n at location y.

$$\begin{aligned} n(y) &= \int_{\mathscr{E}_C}^{+\infty} d\mathscr{E}_y \int_0^{+\infty} d\mathscr{E}_{xz} \left[g_y(\mathscr{E}_y, y)\ g_{xz}(\mathscr{E}_{xz})\ f_{FD}(\mathscr{E}_y + \mathscr{E}_{xz}, \mathscr{E}_F) \right] \\ &= \int_{\mathscr{E}_C}^{+\infty} d\mathscr{E}_y \int_0^{+\infty} d\mathscr{E}_{xz} \left[\sum_{\substack{m \\ \mathscr{E}_m > \mathscr{E}_C}} |\Psi_m(y)|^2\, \delta(\mathscr{E}_y - \mathscr{E}_m)\ \frac{m_D^*}{\pi\hbar^2}\ \frac{1}{1+\exp\left(\frac{\mathscr{E}_y + \mathscr{E}_{xz} - \mathscr{E}_F}{kT}\right)} \right] \\ &= \frac{m_D^*}{\pi\hbar^2} \sum_{\substack{m \\ \mathscr{E}_m > \mathscr{E}_C}} |\Psi_m(y)|^2 \int_{\mathscr{E}_C}^{+\infty} d\mathscr{E}_y \delta(\mathscr{E}_y - \mathscr{E}_m) \int_0^{+\infty} d\mathscr{E}_{xz} \frac{1}{1+\exp\left(\frac{\mathscr{E}_y + \mathscr{E}_{xz} - \mathscr{E}_F}{kT}\right)} \\ &= \frac{m_D^*}{\pi\hbar^2} \sum_{\substack{m \\ \mathscr{E}_m > \mathscr{E}_C}} |\Psi_m(y)|^2 \int_0^{+\infty} \frac{d\mathscr{E}_{xz}}{1+\exp\left(\frac{\mathscr{E}_m + \mathscr{E}_{xz} - \mathscr{E}_F}{kT}\right)} \end{aligned} \tag{6.25}$$

The last integral is the Fermi-Dirac integral with a closed-form solution [115]

$$\int_0^{+\infty} \frac{du}{1+\exp(u-a)} = \ln(1+e^a)\ \ (a \in \mathbb{R}) \tag{6.26}$$

Therefore

$$\int_0^{+\infty} \frac{d\mathscr{E}_{xz}}{1+\exp\left(\frac{\mathscr{E}_m + \mathscr{E}_{xz} - \mathscr{E}_F}{kT}\right)} = kT \ln\left[1+\exp\left(\frac{\mathscr{E}_F - \mathscr{E}_m}{kT}\right)\right] \tag{6.27}$$

Before arriving at the final expression for $n(y)$, it is worth commenting at the result in Equation 6.25:

1. The energy components $\mathscr{E}_y$ and $\mathscr{E}_{xz}$ have different ranges. The potential well requires $\mathscr{E}_y \geq \mathscr{E}_C$, since the bound state energies can only be above the conduction band bottom. On the other hand, $\mathscr{E}_{xz}$ is related to the two directions not restricted by the potential well; therefore, it only needs to be non-negative $\mathscr{E}_{xz} \geq 0$.

2. We used simplified language on the lower bounds of the integrals and summations, such as $\mathscr{E} > \mathscr{E}_C$ in Equation 6.15 and $\mathscr{E}_m > \mathscr{E}_C$ in Equation 6.21. The term $\mathscr{E}_C$ is the minimum value of the position-dependent potential $V(y)$ in the Si body, so that all allowed discrete states are included in the calculation. (The poly-Si gate electrodes

are treated separately and similarly.) For example, as will be seen in Figure 6.13, some energy states $\mathscr{E}_m$ are lower than $V(y)$ around the center of the body. However, electrons in those states may still appear where $\mathscr{E}_C < \mathscr{E}_m < V(y)$.

3. The upper bounds of the energies should be the conduction band top, and when $\mathscr{E}_y > \Delta\mathscr{E}_C = 3.1\,\text{eV}$, the continuous DOS function should be used. However, practically, the continuous free states are always ignored since the integral in Equation 6.27 can be truncated far earlier than $\mathscr{E} = \mathscr{E}_m + \mathscr{E}_{xz}$ reaches 3.1 eV. In fact, for $T = 303\,\text{K}$, it can be shown that $f_{FD}(\mathscr{E}_y + \mathscr{E}_{xz} = 3.1\,\text{eV}, \mathscr{E}_F = 0\,\text{eV}) < 1 \times 10^{-51}$. The vast majority of electrons do not have enough energy to occupy the states much higher than the conduction band bottom.

4. The eigenvalue problem associated with the matrix-form Schrödinger's equation produces as many eigenvalue-eigenvector pairs as the number of mesh points. In our FinFET configuration, $\Delta y = 0.02\,\text{nm}$, and the total simulation space in 1D is 20 nm. Therefore, there may be up to 1000 non-degenerate eigenvalues in the matrix solution; however, only a few of them are of interest. By examining the shape of the wavefunctions (eigenvectors) and energy levels (eigenvalues), we can identify the bound states in the potential well. As will be seen later, there are only 14 out of the 1000 that are the desired bound states. Also, they are not always the lowest 14 eigenvalues, since the potential in gate terminals (out of the potential well) depends on the applied voltage; thus, the bound states need to be identified by their wavefunction shapes, rather than energy values.

5. The y-direction DOS function g_y contains terms $|\Psi(y)|^2$ or the probability of an electron to be present, required by the quantum-mechanical aspect of the potential well. Therefore, $g_y(\mathscr{E}_y, y)$ can also be thought of as a state occupancy function in additional to the Fermi-Dirac function f_{FD} required by thermal equilibrium (Equation 6.16). However, since its unit appears more similar to a "density" quantity, it

is treated as a DOS function. The two interpretations end up with the same result, as the occupancy probability and DOS are multiplied together to produce the total electron concentration. On the other hand, the following integral may be evaluated, using g_{xz} and g_y:

$$\begin{aligned}
\mathrm{DOS}(\mathscr{E}) &= \int_{\mathscr{E}_C}^{\mathscr{E}} d\mathscr{E}_y \int_{-\infty}^{+\infty} dy \left[g_y(\mathscr{E}_y, y)\right] \left[g_{xz}(\mathscr{E} - \mathscr{E}_y)\right] \\
&= \int_{\mathscr{E}_C}^{\mathscr{E}} d\mathscr{E}_y \int_{-\infty}^{+\infty} dy \left[\sum_{\substack{m \\ \mathscr{E}_m > \mathscr{E}_C}} |\Psi_m(y)|^2 \, \delta(\mathscr{E}_y - \mathscr{E}_m)\right] \left[g_{xz}(\mathscr{E} - \mathscr{E}_y)\right] \\
&= \int_{\mathscr{E}_C}^{\mathscr{E}} d\mathscr{E}_y \left[\sum_{\substack{m \\ \mathscr{E}_m > \mathscr{E}_C}} \delta(\mathscr{E}_y - \mathscr{E}_m)\right] \left[g_{xz}(\mathscr{E} - \mathscr{E}_y)\right] \\
&= \left[\sum_{\substack{m \\ \mathscr{E}_m > \mathscr{E}_C}} U(\mathscr{E} - \mathscr{E}_m)\right] \left[\frac{m_D^*}{\pi\hbar^2}\right] \\
&= \left(\frac{m_D^*}{\pi\hbar^2}\right) \times \begin{pmatrix} \text{\# of } y\text{-direction eigenstates} \\ \text{between CB bottom and } \mathscr{E} \end{pmatrix}
\end{aligned} \tag{6.28}$$

The above quantity has unit $\mathrm{cm}^{-2}\mathrm{eV}^{-1}$, and it represents the area density of states *available for electrons to occupy*. The unit step function is defined as

$$U(\mathscr{E} - \mathscr{E}_m) = \begin{cases} 0 & \mathscr{E} < \mathscr{E}_m \\ 1 & \mathscr{E} \geq \mathscr{E}_m \end{cases} \tag{6.29}$$

Finally, the electron concentration $n(y)$ at location y is found by evaluating the integral in Equation 6.25:

$$n(y) = \frac{m_D^* kT}{\pi\hbar^2} \sum_{\substack{m \\ \mathscr{E}_m > \mathscr{E}_C}} |\Psi_m(y)|^2 \ln\left[1 + \exp\left(\frac{\mathscr{E}_F - \mathscr{E}_m}{kT}\right)\right] \tag{6.30}$$

The boundary conditions are listed in Table 6.4.

Table 6.4: Boundary conditions used at the semiconductor-oxide interfaces for the Schrödinger-Poisson equation system

Unknown Quantity and Expression	Location	Condition	Reason
Potential $\phi(y)$	Simulation boundaries	Equilibrium value minus bias V_{GS}	External force
Wavefunction $\Psi(y)$	Simulation boundaries	Zero	Assuming bound states
Displacement field $\varepsilon(y)\frac{d}{dy}\phi(y)$	Semiconductor-oxide interfaces	Continuous*	No singular surface charge
de Broglie Wave Velocity $\frac{1}{m_M^*(y)}\frac{d}{dy}\Psi(y)$	Semiconductor-oxide interfaces	Continuous*	Continuous potential

* By integrating the Gauss's law or Schrödinger's equation once across the interface.

As described previously in the workflow in Section 6.3.1, we apply finite difference to all quantities and equations, which are solved numerically.

First, the Poisson's equation (6.11) is solved as a non-linear matrix problem with the Newton-Raphson method. The result with $V_{GS} = 0$ serves as the first initial guess when we immediately solve it again with user input $V_{GS} \neq 0$. The second result is the classical solution, which also becomes the second initial guess for quantum correction.

Next, Equations 6.11 (Poisson's equation), 6.17 (Schrödinger's equation), 6.18 (mesoscopic potential), 6.30 (electron concentration of the bound states), and 6.14 (hole concentration following the law of mass action) are combined and become a set of simultaneous equations. They are solved in a self-consistent loop that generally follows the gradient in the search space, but no derivatives are evaluated.

The Schrödinger's equation is solved as an eigenvalue problem, while the Poisson's equation is solved as a linear matrix problem. Note that in our second step applying quantum correction, the right-hand side of the Poisson's equation, containing the charge density, is known prior to the solution; therefore, the corresponding matrix equation can be solved using linear algebra techniques, including direct methods (e.g., matrix inversion,

Gaussian elimination, Cholesky decomposition[25]) and iterative methods. No matter what technique is used, both the left-hand-side matrix and right-hand-side vector in the equation are constant and do not depend on the unknown variable (potential as a vector).

These two equations are solved separately in repeated cycles. In each iteration, the Schrödinger's equation is solved first for energy eigenvalues $\mathscr{E}_m$ $(m = 1,2,3,\dots)$ and wavefunctions $\Psi_m(y)$ using the latest potential solution $V(y)$. The charge density $\rho(y)$ is updated using the latest Schrödinger's equation solutions, and then the Poisson's equation is solved for electrostatic potential $\phi(y)$. The mesoscopic potential $V(y)$ is updated with the latest $\phi(y)$, and the Schrödinger's equation is solved again.

This process repeats until the absolute and relative changes to all quantities solved for reach below a pre-defined threshold. Although convergence to a consistent set of solution quantities is guaranteed for a physical system under any reasonable applied voltage, since the potential and wavefunction are calculated separately, there is no general guarantee that the loop of numerical calculations will converge with the initial guess from the classical solution; in fact, prior attempts without the following improvement end up with diverging results after only a few iterations.

The main challenge lies in the Poisson's equation in the iteration process. Because directly solving the finite-difference matrix equation is effectively finding $\phi(y)$ by integrating the Poisson's equation twice, a small numerical error at a local mesh point may propagate and cause significant global discrepancies. The calculation can crash due to unbounded errors if the new potential or wavefunction solution from an iteration step is far from physically reasonable.

To assist convergence, the changes to the electrostatic potential solution $\Delta\phi(y)$ is empirically and substantially reduced before evaluating the wavefunction in the next iteration, at the price of more iteration steps. In other words, the iteration process is effectively a damped gradient descent method.

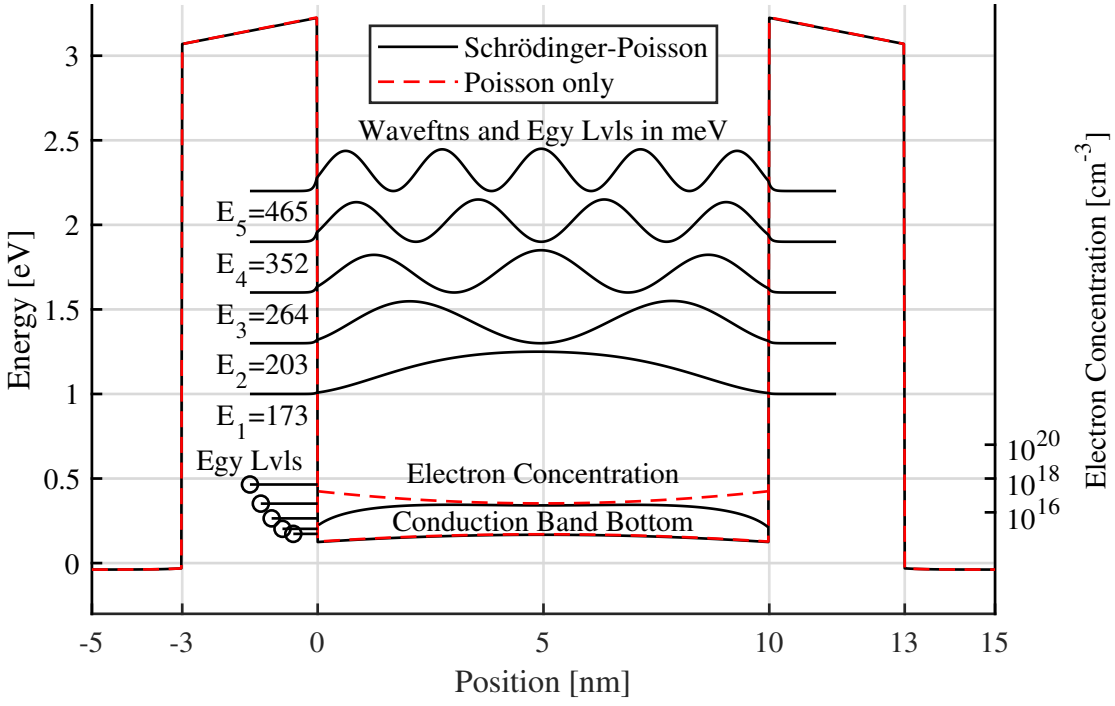
[25]The finite-difference matrix is real symmetric, and therefore Hermitian; using the Cholesky method is the ideal way *and the default way* implemented by Matlab's built-in matrix solver.

6.3.3 Simulation Results

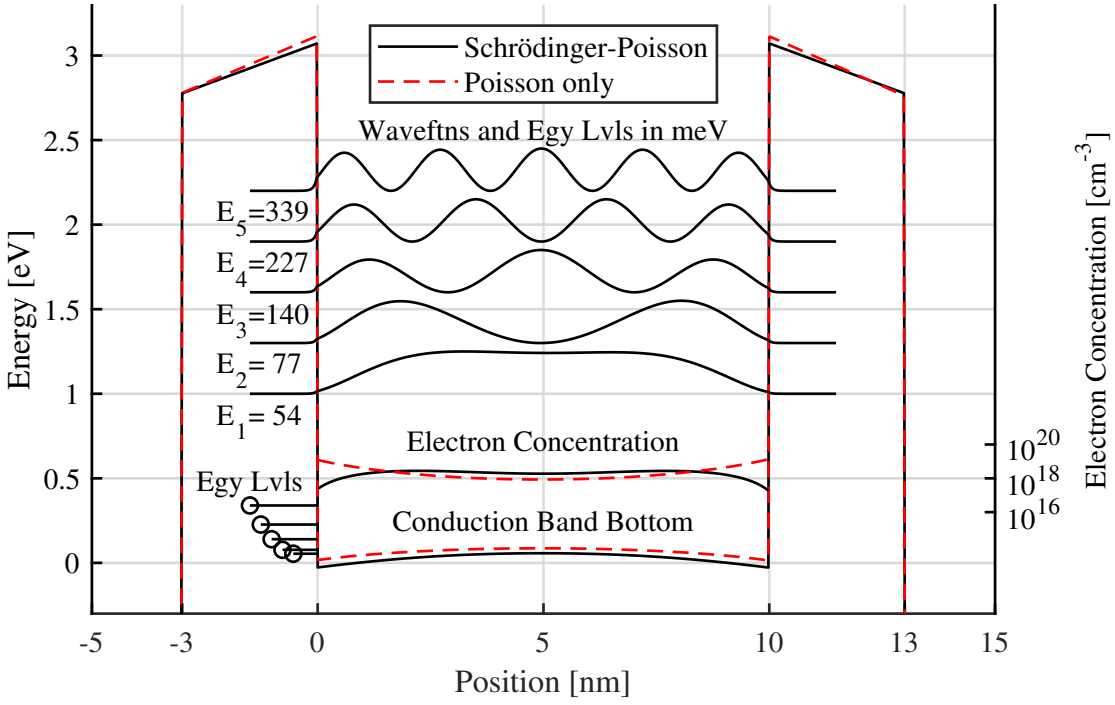
Selected calculation results to the 1D potential well problem inside the same FinFET as previous are shown in Figures 6.13-6.15. In Figure 6.13, the position (y)-dependent data under zero and positive bias voltages is plotted, including:

- Mesoscopic potential $V(y)$ in eV, which is also the band diagram of the conduction band bottom.
- Normalized wavefunction square-magnitude $|\Psi_m(y)|^2$ for the first five eigenstates ($m = 1,2,3,4,5$) offset at arbitrary locations (although they always quickly cease to zero once entering the oxide regions), and their —
- Associated energy eigenvalues $\mathscr{E}_m$ in meV and their true positions in the potential well (match sticks near the bottom).
- Electron concentration $n(y)$ (counting only bound states) in cm^{-3}.

As mentioned before, the Fermi level of the p-type body is set to zero ($\mathscr{E}_F = 0$). The quantum mechanical results (solid lines) are overlayed with the classical solutions (red dashed lines) for $V(y)$ and $n(y)$.

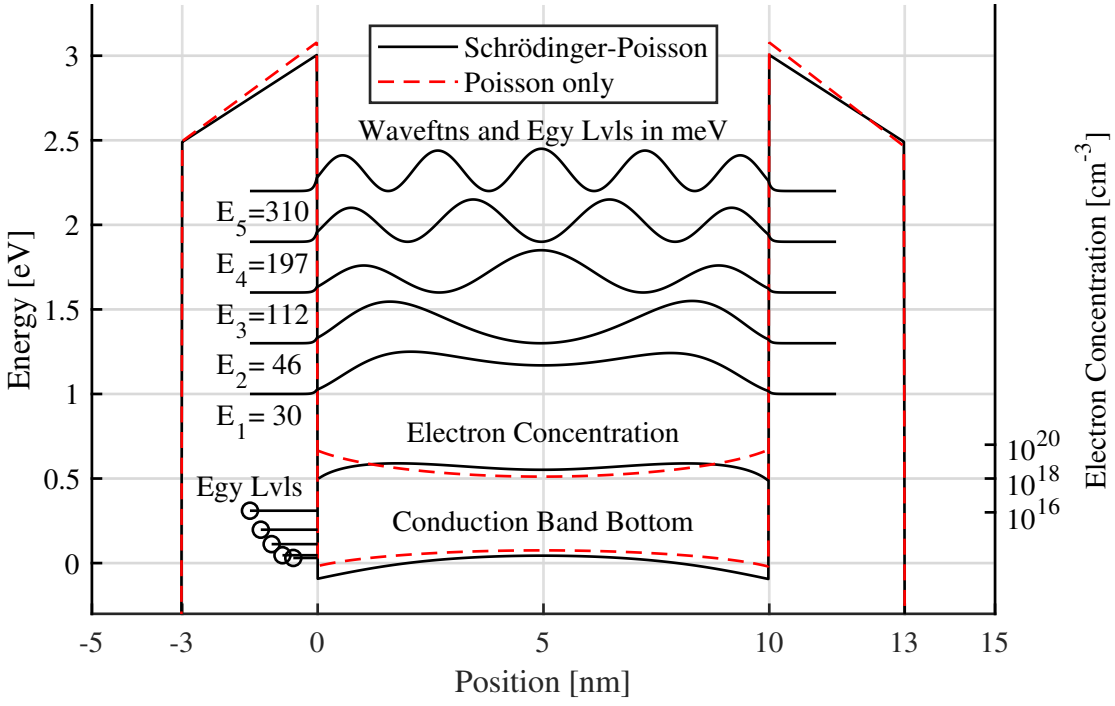


(a) $V_{GS} = 0.0\,\mathrm{V}$

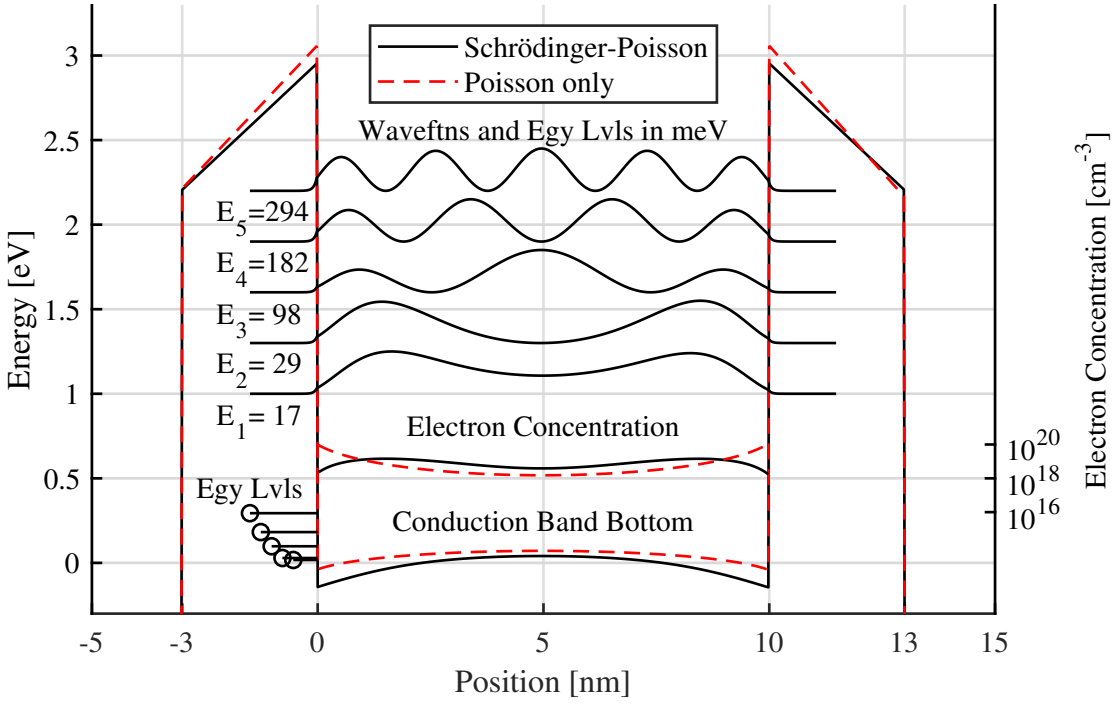


(b) $V_{GS} = 0.3\,\mathrm{V}$

Figure 6.13: The 1D quantum corrected solution to the studied FinFET. In this page: $V_{GS} = 0.0\,\mathrm{V}$ (a) and $0.3\,\mathrm{V}$ (b). In these plots, position-dependent data under several bias voltages are shown. More details are in the text.



(c) $V_{GS} = 0.6\,\mathrm{V}$



(d) $V_{GS} = 0.9\,\mathrm{V}$

Figure 6.13: (Continued) The 1D quantum corrected solution to the studied FinFET. In this page: $V_{GS} = 0.6\,\mathrm{V}$ (c) and $0.9\,\mathrm{V}$ (d). In these plots, position-dependent data under several bias voltages are shown. More details are in the text.

From the results, a few key points can be observed. The wavefunctions $\Psi_m(y)$ (not shown) confirm that the fin structure mainly resembles a square potential well, and $|\Psi_m(y)|^2$ of the first five eigenstates ($m = 1,2,3,4,5$) are similar to the standing-wave solution to the ideal infinite-barrier square potential well, which have integer numbers of sinusoidal waves. Thus, $|\Psi_m(y)|^2$ or conceptually the probability of an electron's existence in the eigenstate m at location y generally decreases approaching the interfaces, and exponentially decays once entering the insulating dielectric regions (SiO_2). As a result, the total electron concentration $n(y)$ are lower at the interfaces than further into the body, or effectively, the electrons are "pushed" into the body. This contrasts with the classical results, which suggest the highest $n(y)$ should be at the interfaces. At $V_{GS} = 0.9\,\mathrm{V}$ for example, $n(y = 0\,\mathrm{nm}) = 2.3 \times 10^{18}\,\mathrm{cm}^{-3}$ at the interface from the quantum-corrected solution, whereas $n(y = 5\,\mathrm{nm}) = 3.9 \times 10^{18}\,\mathrm{cm}^{-3}$ in the middle of the body. From the classical solution, the readings at the two positions are $9.2 \times 10^{19}\,\mathrm{cm}^{-3}$ and $1.5 \times 10^{18}\,\mathrm{cm}^{-3}$, respectively.

Additionally, the potential is different between the two solutions. An apparent reason is $\phi(y)$ and $n(y)$ directly depend on each other. Also in the quantum-corrected solution, $n(y)$ is lower right at the surfaces, so the interface E field is smaller, according to Gauss's law. The oxide E field is also smaller due to boundary conditions (continuity of perpendicular displacement field), which calls for less band bending in the oxide. The oxide E field will be further discussed in the next section (Section 6.3.4). However, the barrier heights right at the body-oxide and gate electrode-oxide interfaces remain unchanged.

Besides the square potential well solutions (sinusoidal wave-like wavefunctions), one can observe that as the bias V_{GS} increases, the solutions progressively migrate to triangular potential well ones (Airy function-like wavefunctions). In Figure 6.13, this is most obvious for the first state $|\Psi_1(y)|^2$, as $\mathscr{E}_1$ gradually becomes lower than the highest potential $V(y)$ inside the body. This eigenstate, bound by the oxide-body-oxide potential well, transitions to be bound by the oxide-body triangular well. At $V_{GS} = 0.9\,\mathrm{V}$, two peaks in $|\Psi_1(y)|^2$ are present, each related to the Airy function-like solution at each interface.

A clearer sight is given in Figure 6.14. Inside the oxide-body-oxide well from the FinFET structure, there are fourteen (sometimes fifteen[26]) bound states with their energies $\mathscr{E}_m$ ($m = 1, 2, \ldots, 14$, black solid lines) lower than the oxide barrier height $\Delta\mathscr{E}_C = 3.1\,\text{eV}$ (dash-dotted line on top).

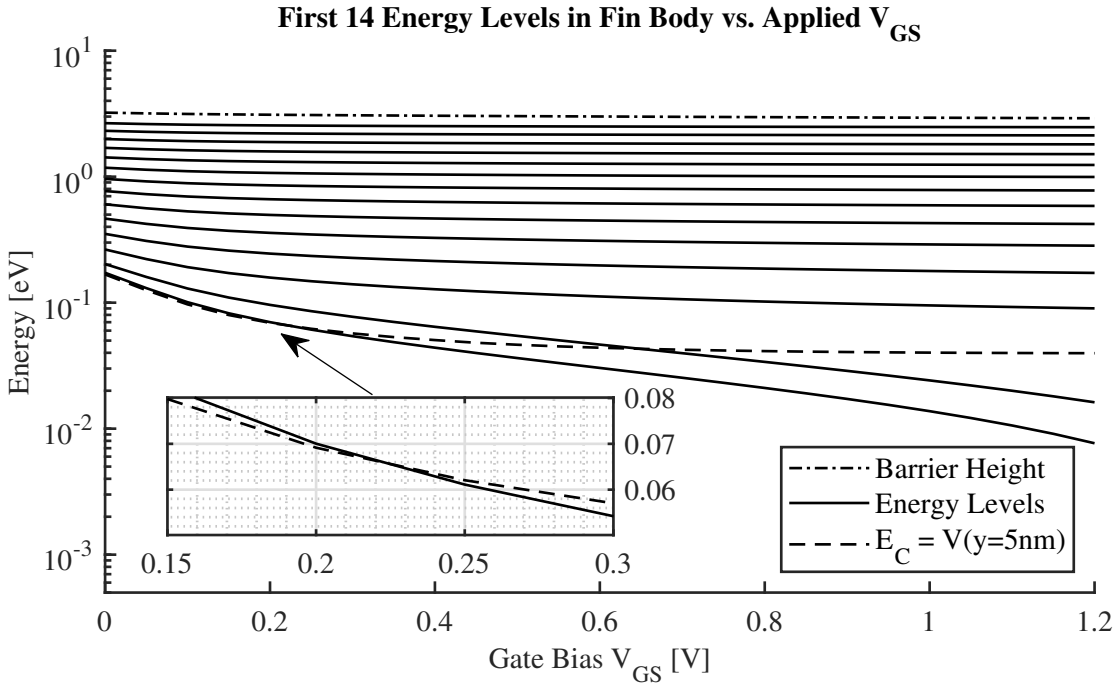


Figure 6.14: First (lowest) fourteen energy eigenvalues $\mathscr{E}_m$ ($m = 1, 2, \ldots, 14$, solid lines from bottom to top in order) in eV versus applied gate bias V_{GS} in the FinFET 1D quantum-corrected solution. The barrier height $\Delta\mathscr{E}_C = 3.1\,\text{eV}$ (dash-dotted line on top) is the change in the conduction band bottom energy across the body-oxide interface. The peak potential (dashed line on bottom) is the peak body potential $V(y = 5\,\text{nm})$.

As V_{GS} increases, the first eigenstate's energy $\mathscr{E}_1$ goes from slightly (less than 10 meV) higher than the peak potential $V(y)$ inside the body (which is at the center,

[26]Fourteen states in total with the highest energy eigenvalue $\mathscr{E}_{14} \approx 2.5\,\text{eV}$ are always bound states. All are included in favor of visual appearance. The fifteenth eigenstate's energy $\mathscr{E}_{15}$ is very close to the barrier height and sometimes exceeds it (when it becomes a free state). However, due to numerical accuracy limits, the change in $\mathscr{E}_{15}$ versus V_{GS} does not have a consistent trend, and it is hard to improve the errorneous results. Furthermore, because the wavefunctions are forced to be zero at simulation boundaries (Dirichlet boundary conditions), the "free" states calculated in the simulation does not represent the true traveling waves. Nevertheless, only the first several (three to five) bound states are used to evaluate the total electron concentration. In fact, the Fermi-Dirac occupancy probability of the 14th state is $f = \left[1 + \exp\left(\frac{2.5\,\text{eV}}{kT}\right)\right]^{-1} \approx 1 \times 10^{-42}$, while the additive relative precision of a modern computer (64-bit double precision floating number) is $\approx 1 \times 10^{-16}$.

$y = 5\,\text{nm}$ because of geometric symmetry) to lower than that around $V_{GS} = 0.2$–$0.3\,\text{V}$, in accordance to the transition from Figure 6.13a to 6.13b. The second state also goes through this transition around $V_{GS} = 0.6$–$0.7\,\text{V}$, which is less apparent in Figure 6.13 but still observable.

Overall, the total electron concentration $n(y)$ increases as the forward bias V_{GS} increases. This is confirmed by finding the area electron concentration n_s in cm^{-2}, or effectively the total number of electrons under inversion per gate area. One may want to keep in mind that since the simulated FinFET has two oxide-semiconductor interfaces, effectively it has two gates in the traditional sense, and the calculation of n_s always includes both. For the 1D, quantum-corrected and classical solutions,

$$n_{s,1D} = \int_{y=0\,\text{nm}}^{y=10\,\text{nm}} n(y)\,dy \tag{6.31}$$

The "area" dimensions are the fin height ($T_{BODY} = 34\,\text{nm}$, z direction) and the channel length ($L_G = 25\,\text{nm}$, x direction) which are not simulated. The results under bias range of $V_{GS} = 0$–$1.2\,\text{V}$ are plotted in Figure 6.15.

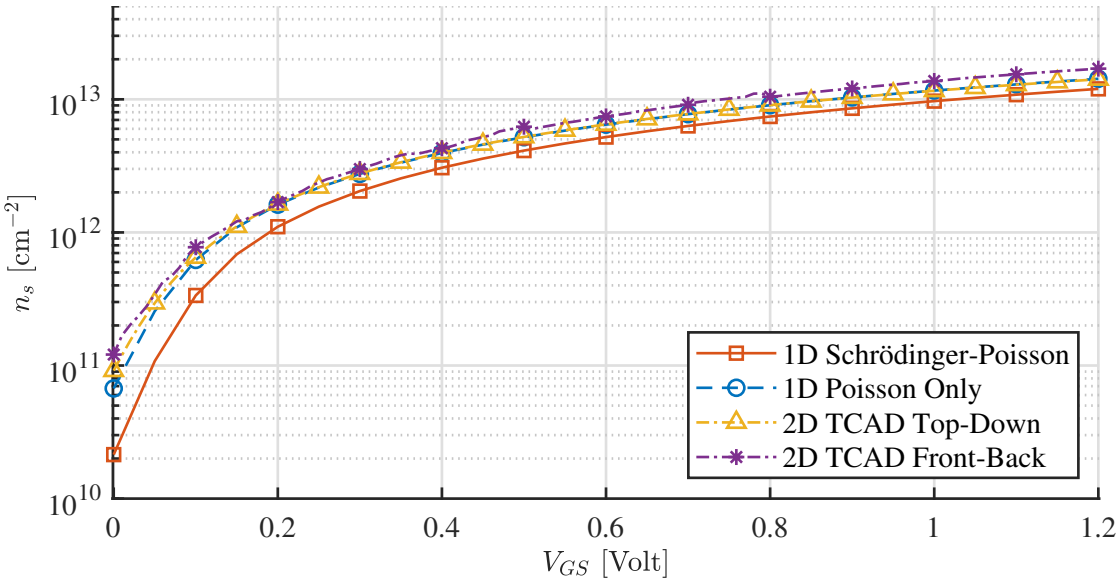


Figure 6.15: Area electron concentration $n_s\,[\text{cm}^{-2}]$ versus applied gate bias V_{GS} in the FinFET. Included results are from the 1D quantum-corrected solution (Schrödinger's and Poisson's equations), 1D classical solution (Poisson's equation only), and 2D Cider (TCAD) solutions from top-down ($V_{DS} = 0\,\text{V}$) and front-back views.

For comparison, the results from the 2D top-down view ($V_{DS} = 0\,\mathrm{V}$) and front-back view simulations are also included. For the top-down view, the line integral is taken at the center of the channel, so

$$n_{s,2D}\ (\text{top-down}) = \int_{y=0\,\mathrm{nm}}^{y=10\,\mathrm{nm}} n\,(x = 12.5\,\mathrm{nm}, y)\,dy \tag{6.32}$$

For the front-back view, the following spatial average is evaluated. The result is exactly the same as in Figure 6.8a.

$$n_{s,2D}\ (\text{front-back}) = \frac{1}{34\,\mathrm{nm}} \int_{z=0\,\mathrm{nm}}^{z=34\,\mathrm{nm}} \int_{y=0\,\mathrm{nm}}^{y=10\,\mathrm{nm}} n\,(y,z)\,dydz \tag{6.33}$$

We can observe that the p-type body, which is already partially inverted at zero bias, becomes more abundant of electrons as V_{GS} increases; for the quantum-corrected results, n_s is more than 10 times higher at $V_{GS} = 0.1\,\mathrm{V}$ and more than 100 times higher at $V_{GS} = 0.5\,\mathrm{V}$ than at $V_{GS} = 0\,\mathrm{V}$. Furthermore, the quantum-corrected calculation reports generally lower concentration than the classical calculation. At $V_{GS} = 0\,\mathrm{V}$, it is $2.1 \times 10^{10}\,\mathrm{cm}^{-2}$ (quantum) compared to $6.7 \times 10^{10}\,\mathrm{cm}^{-2}$ (classical), and $1.2 \times 10^{13}\,\mathrm{cm}^{-2}$ at $V_{GS} = 1.2\,\mathrm{V}$ compared to $1.4 \times 10^{13}\,\mathrm{cm}^{-2}$.

Meanwhile, the 1D classical solution is almost identical to the 2D top-down view solution from Cider, with the exceptions at low bias; at $V_{GS} = 0\,\mathrm{V}$, the top-down simulation yields $9.2 \times 10^{10}\,\mathrm{cm}^{-2}$ at the center of the body. It could be because in the top-down view, the drain and source dopants help with the channel inversion by depleting the body, which is categorized as the short-channel effect or the charge-sharing effect [99], as mentioned in Section 6.1. For $V_{GS} > 0.2\,\mathrm{V}$, the difference is minimal, less than 1 %. The 2D front-back simulation using Cider generally reports higher results: $1.2 \times 10^{11}\,\mathrm{cm}^{-2}$ at $V_{GS} = 0\,\mathrm{V}$ and $1.7 \times 10^{13}\,\mathrm{cm}^{-2}$ at $V_{GS} = 1.2\,\mathrm{V}$. The additional gate structure on the top helps inverting the body.

Another quantum effect is the presence of non-valence-band electrons in the oxide region. The material SiO_2 is usually considered as an insulator, as its bandgap ($\approx 9\,\mathrm{eV}$) is too large for any valence-band electrons to become thermally activated and populate the

conduction band. In a SiO_2-Si structure, the Si conduction band bottom is aligned to around the middle of the bandgap in SiO_2. Therefore from classical mechanics, it is impossible for Si conduction-band electrons to move into SiO_2 across the interface. However, quantum mechanical solutions reveal that a substantial amount of electrons are present in the SiO_2 region.

The band structure and first three wavefunctions at $V_{GS} = 0.9\,\mathrm{V}$, identical to Figure 6.13d, are shown again in Figure 6.16. The real parts of the wavefunctions $\Re\{\Psi_m\}$ are drawn for $m = 1,2,3$, offset by arbitrary distances indicated by the dashed lines.

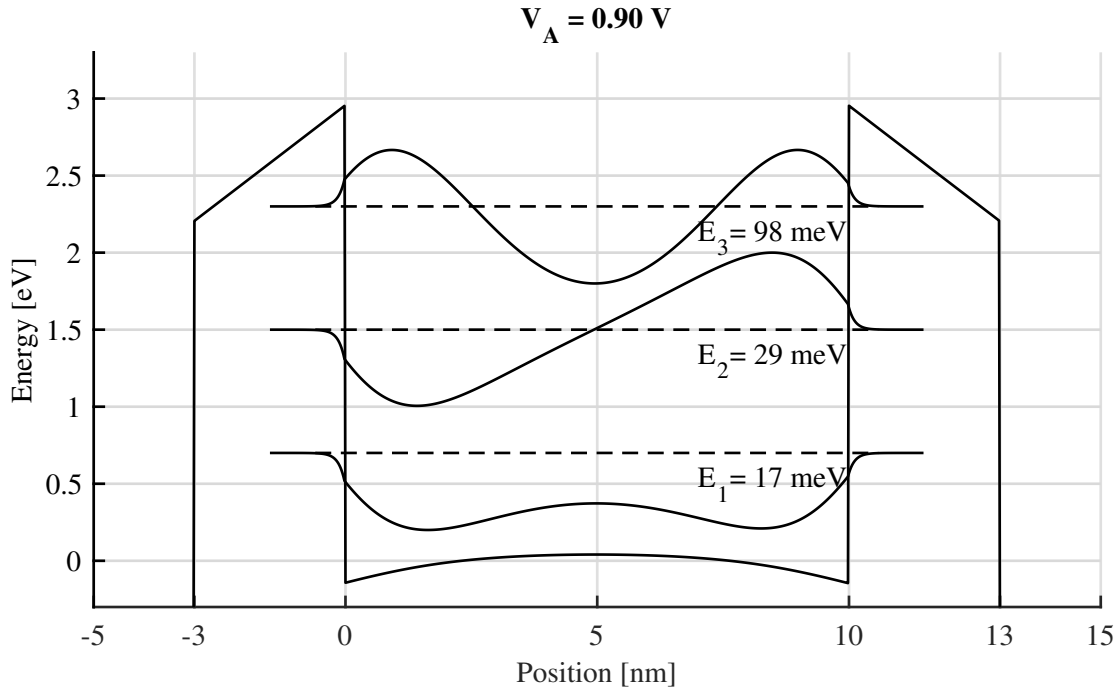


Figure 6.16: Quantum-corrected potential $V(y)$ and wavefunctions (real part) $\Re\{\Psi_m\}$, $m = 1,2,3$ at $V_{GS} = 0.9\,\mathrm{V}$.

In the fin body, because of the symmetric potential barrier, the wavefunctions Ψ_m are symmetric standing waves (ideally taking real values). In the gate oxides, they become evanescent waves. The exponential spatial decay rate can be estimated by a rectangular barrier solution[27]

[27] Using the WKB approximation to a triangular barrier solution is more accurate but more complicated. It is avoided here.

$$\Psi_m \approx A_m e^{-\kappa_m y} \ (-3\,\text{nm} < y < 0\,\text{nm},\ m = 1,2,3,\dots) \tag{6.34}$$

where A_m are normalization coefficients, and the real wavenumbers κ_m are given by

$$\kappa_m = \sqrt{\frac{2m_M^*\left(\Delta\mathscr{E}_C - \mathscr{E}_m\right)}{\hbar^2}} \tag{6.35}$$

which are $8.338\times10^9\ \text{m}^{-1}$, $8.322\times10^9\ \text{m}^{-1}$, and $8.228\times10^9\ \text{m}^{-1}$ for $m = 1,2,3$, respectively, or $\kappa_m^{-1} \approx 0.12\,\text{nm}$. The evanescent waves suggest that the electrons from the Si conduction band may appear in the oxide region as far as 1 nm away from the interface. However, before being possibly scattered into the SiO_2 conduction band by phonons, they remain in the SiO_2 bandgap, which normally forbids electrons from existing. Therefore, these electrons "tunneling" or "penetrating" into the oxide are highly localized and do not contribute directly to channel conduction, although defect states and phonon scattering may help them form a secondary channel in parallel to the one in Si [116]. Additionally, these electrons can be trapped by interface states and affect the device's characteristics such as threshold voltage (ΔV_{TH}) and subthreshold swing (SS). They may even assist the formation of defect sites that eventually can contribute to the tunneling current [42].

To quantitatively describe the electron penetration effect, we calculate the "surface charge" or per-area electron concentration n_s in the oxide, i.e.,

$$n_{s,\text{oxide}} = \int_{y=-3\,\text{nm}}^{y=0\,\text{nm}} n(y)\,dy + \int_{y=10\,\text{nm}}^{y=13\,\text{nm}} n(y)\,dy \tag{6.36}$$

Note that in practical calculations based on the numerical, finite-difference solutions, $n(y)$ at the very points of material interfaces (e.g., y=-3 nm, y=0 nm, etc.) may represent the Si conduction band electrons depending on the particular mesh configuration, and thus may need to be excluded from the summation for $n_{s,\text{oxide}}$. The results are shown in Figure 6.17 versus V_{GS}. Included are: $n_{s,\text{body}}$ from the 1D quantum corrected solution (identical to Equation 6.31 and Figure 6.15), oxide electron concentration $n_{s,\text{oxide}}$ using Equation 6.36, the sum ($n_{s,\text{oxide}} + n_{s,\text{body}}$), and the ratio ($n_{s,\text{oxide}}/n_{s,\text{body}}$). The 1D classical (Poisson's equation only) solution, identical to Figure 6.15, is also included for comparison. The

electrons in the oxide make up for 3.47 % of the total electrons (in body and oxide) at $V_{GS} = 0\,\text{V}$ and 12.8 % at $V_{GS} = 1.2\,\text{V}$.

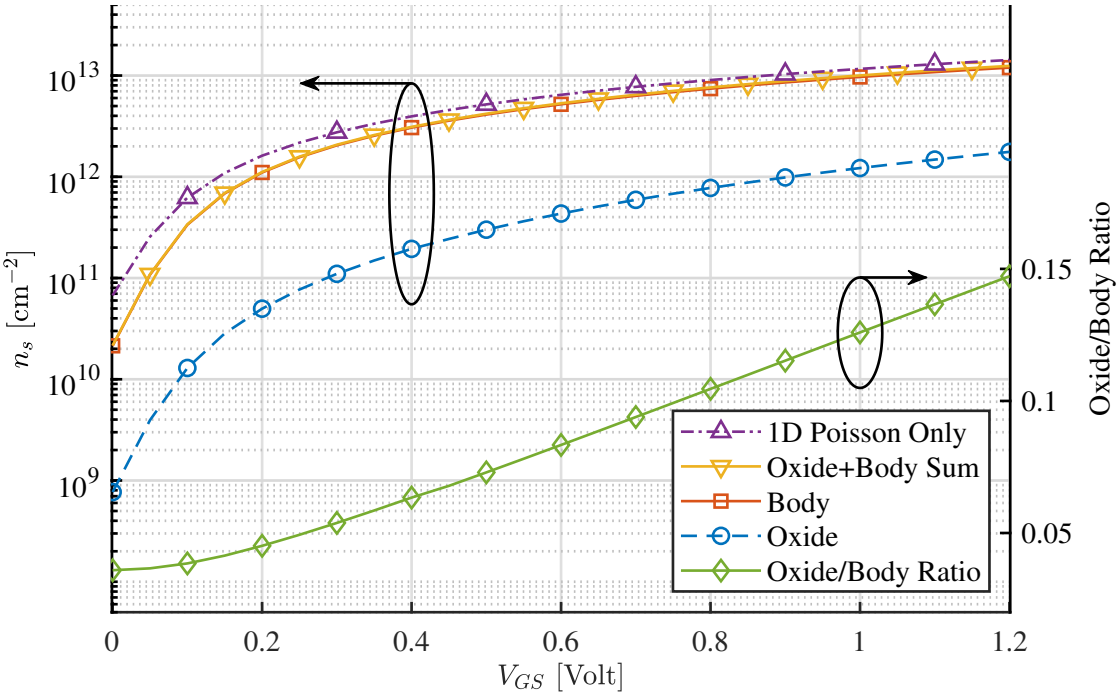


Figure 6.17: Electron area concentration n_s in the fin body (Si), the gate dielectric (SiO_2), and the classical solution in the body for comparison. All results are from the 1D solutions.

To summarize, the quantum-corrected calculation of the investigated FinFET generally results in lower surface electron concentrations than the classical calculations, and electrons are "pushed" into the body away from the interfaces. However, when considering the channel current, one may need to take into account the change in mobility due to less surface scatterings. Meanwhile, electrons from the Si body can penetrate into the SiO_2 bandgap of the gate dielectric regions, leading to potential additional problems in the device characteristics and reliability.

The quantum calculation also results in generally lower oxide fields, which will be further discussed next, in Section 6.3.4. A closed form expression of the oxide field $E_{OX}\,(V_{GS})$ is extracted and used for the oxide breakdown circuit model.

6.3.4 Extracting the FinFET Gate Oxide Field

In Section 6.4, the gate dielectric breakdown model requires calculating the oxide field, given applied voltage V_{GS}. This section analyzes E_{OX}, the electric (E) field in the simulated FinFET's gate dielectric (SiO_2).

In the 1D simulations (quantum-corrected Schrödinger's and Poisson's equations, and classical Poisson's equation only), the E field is always in the perpendicular or the y direction. Therefore, $E_{OX} = E = \pm\frac{d\phi}{dy}$ in 1D is calculated for each V_{GS}. (The polarity depends on the geometrical arrangement.)

In the 2D simulation (Cider, classical Poisson's equation), the E field could rise between the gate and body (mainly from V_{GS}) and between the gate and drain (mostly from V_{DS}). Although they need to be calculated differently, the relations between the gate-drain field and V_{GD} and between the gate-body field and V_{GS} are similar. Therefore, in this section, only the perpendicular field $E_{OX} = E_\perp$ depending on V_{GS} is analyzed with $V_{DS} = 0\,\mathrm{V}$. The arithmetic average of E_{OX} in all oxide regions is taken for each V_{GS}. The dependency on V_{DS} can always be modeled separately following the same workflow.

In the top-down view, $E_\perp = \pm E_y$. The field is sampled at the center of the body ($x = 12.5\,\mathrm{nm}$) and inside the oxide regions ($-3\,\mathrm{nm} \le y \le 0\,\mathrm{nm}$ and $10\,\mathrm{nm} \le y \le 13\,\mathrm{nm}$). In the front-back view, the sampled regions are —

1. $34\,\mathrm{nm} < z < 37\,\mathrm{nm},\ 1\,\mathrm{nm} < y < 9\,\mathrm{nm}$ (top gate, $E_\perp = -E_z$),
2. $5\,\mathrm{nm} < z < 30\,\mathrm{nm},\ -3\,\mathrm{nm} < y < 0\,\mathrm{nm}$ (left gate, $E_\perp = E_y$), and
3. $5\,\mathrm{nm} < z < 30\,\mathrm{nm},\ 10\,\mathrm{nm} < y < 13\,\mathrm{nm}$ (right gate, $E_\perp = -E_y$).

The four corners are ignored as the finite difference method using rectangular mesh grids results in more than $\pm 10\,\%$ ($V_{GS} = 1.0\,\mathrm{V}$) of difference from the rest regions, and counting such fluctuations may be unnecessarily over-specific since the fringing effect highly depends on the fabricated device geometry. These geometric corners can always be modeled separately following the same workflow.

The extracted oxide perpendicular field E_{OX} is shown in Figure 6.18 versus applied gate bias V_{GS}.

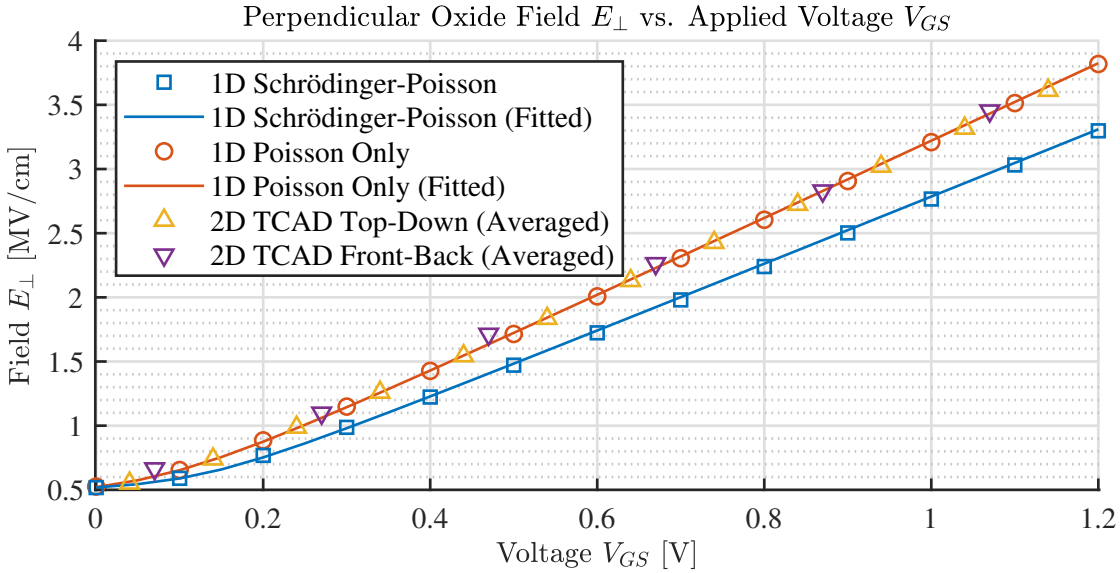


Figure 6.18: Extracted gate oxide field E_{OX} [MV/cm] (perpendicular components only) versus applied gate bias V_{GS} from various simulations and fitted closed-form expressions for selected cases.

The field-voltage relationship is almost linear, except for the near-zero voltages. It can be shown that for negative and decreasing V_{GS} (or more generally, $V_{GS} < V_{TH}$), there is a "flat" region where E_{OX} stays relatively constant and the body changes from inversion to accumulation, and then E_{OX} continues to decrease linearly (even $E_{OX} < 0$ is possible) as the accumulation of holes in the p-type body continues to build up.

Next, a compact model of the FinFET oxide field is proposed and extracted using data from the 1D quantum-corrected and the 2D classical (top-down) simulations. The most straightforward linear relation is for a parallel-plate capacitor of thickness t, where the dielectric field E under applied voltage V is $E = V/t$. In a semiconductor device where the electrons in the body are not described as "sheet charge" or a singularity right at the metallic electrode, additional terms are needed to describe the near-linear relationship. In order to accommodate the different slope and intercept, two coefficients are added. Since in the oxide breakdown model, only non-negative V_{GS} is considered, for simplicity, it is

assumed that E_{OX} asymptotically reaches a minimum value as V_{GS} decreases. To limit E_{OX} at low voltages above such minimum, a quadratic smoothing function similar to Equation 2.39 is used, introducing two additional fitting parameters. The resulting model expressions for $E_{OX}\,(V_{GS})$ are

$$E_{OX} = \frac{1}{2}\left(E'_{OX} + E_O + \delta E_O + \sqrt{\left(E'_{OX} - E_O + \delta E_O\right)^2 + 4E_O\ \delta E_O}\right) \tag{6.37a}$$

$$E'_{OX} = \frac{V_{GS} + V_O}{A\ t_{OX}} \tag{6.37b}$$

where $t_{OX} = 3\,\text{nm}$. The set of four parameters E_O, δE_O, V_O, and A are extracted using the in-house Genetic Algorithm from the 1D quantum-corrected solutions and the 1D classical solutions, calculated in Section 6.3. A total of 4×10^6 trials are evaluated. The parameters are listed in Table 6.5. The evaluated expressions are plotted in Figure 6.18 together with other simulation data.

Table 6.5: Extracted parameters of the FinFET gate oxide field model from 1D simulation. (Classical: Poisson's equation only; Quantum: Schrödinger's and Poisson's equations.)

Symbol	Purpose	Unit	Value (Classical)	Value (Quantum)
V_O	Voltage offset	mV	14.67	33.80
A	Linear slope	—	1.099	1.267
E_O	Lower limit	kV/cm	383.3	452.2
δE_O	Limit smoothing	kV/cm	126.1	54.93

From the extracted model parameters, several observations can be summarized. Naturally, according to the textbook knowledge of the threshold voltage V_{TH}, the gate oxide field is

$$E_{OX} = \frac{V_{GS} - V_{TH}}{t_{OX}} \tag{6.38}$$

However, the extracted intercepts (V_O) are different than $V_{TH} \approx 0.1\,\text{V}$ extracted in Section 6.2, suggesting the surface potential in the FinFET be different than in a traditional planar device. Also, the slope A is $\sim 15\,\%$ larger with quantum correction than the classical solution. A phenomenological interpretation is that the FinFET gate dielectric is effectively thicker, and the physical reason is that the channel electrons are generally "pushed" into

the body, so about 1–2 nm of the body (Si) near the interface can contribute as additional dielectric (SiO_2). Alternatively, one may rewrite Equation 6.37b as

$$\begin{aligned} E'_{OX} &= \frac{A^{-1}V_{GS} + A^{-1}V_O}{t_{OX}} \\ &= \frac{V_{GS} - \left[\left(1 - A^{-1}\right)V_{GS} - A^{-1}V_O\right]}{t_{OX}} \end{aligned} \tag{6.39}$$

Compared to Equation 6.38, the constant term V_{TH} becomes a function of V_{GS}, suggesting another interpretation that the surface potential under inversion has a stronger dependency on V_{GS} in the FinFET than it does in a traditional planar device. At last, the lower limit of E field involves a potential difference of 1.15 V (classical) or 1.36 V (quantum). For reference, the Fermi level mismatch (before band bending) between the n++-type gate ($\sim \frac{1}{2}\mathscr{E}_g = 0.55\,\text{eV}$) and the p-type substrate ($\phi_p = V_T \ln \frac{N_A}{n_i} = -0.50\,\text{V}$) is 1.05 V.

6.4 Empirical Circuit Model of Rapid Gate Oxide Breakdown

Based on the device-level simulation data and gate oxide rapid breakdown (RBD) experiment data, one can predict the condition of a transistor's gate oxide (whether or not it has experienced a rapid and permanent BD event) from its bias voltages. Specifically, we project our knowledge about "thick" oxides in planar devices, which we experimentally tested in Section 5.4, to the FinFET devices studied so far with ultra-thin SiO_2 gate dielectrics, using the compact E field model $E_{OX}(V_{GS})$ extracted in Section 6.3.4.

The goal is to evaluate the RBD condition together with transient (time-dependent), electronic (circuit-level) simulations. Therefore, a SPICE model is created as follows.

The overall equivalent circuit contains a single shunt resistor R_G between the transistor M1's gate and source terminals, as illustrated in Figure 6.19. As mentioned before, in this work, only the V_{GS} dependency of the oxide field E_{OX} is analyzed, while one can always add similar models for V_{DS} dependency separately.

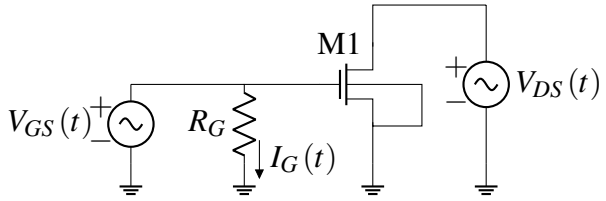


Figure 6.19: Equivalent SPICE circuit for the gate oxide RBD model. The resistor R_G represents the dielectric's always-existing leakage current and the conductive path after RBD. M1 is a regular transistor model without the DC gate current calculation.

The resistance R_G depends on the BD condition (yes or no). Before RBD, the oxide is mainly insulating, so $R_G \gg 1\,\mathrm{M\Omega}$. When the gate field becomes higher than a threshold, RBD happens, and the gate oxide becomes broken when R_G reduces to the measured post-RBD value. R_G maintains at such low value once the gate is broken, no matter how V_{GS} fluctuates. This information needs to be kept in a "memory" as the device's RBD state. An RC circuit is built for this purpose, as shown in Figure 6.20.

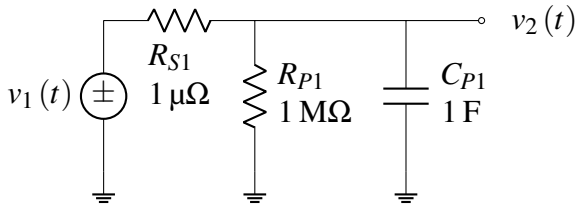


Figure 6.20: Memory circuit to store the oxide RBD state.

The capacitor C_{P1} stores the maximum gate field the device has ever experienced as of simulation time t (1 V across C_{P1} means 1 MV/cm of field across the oxide), or

$$\begin{aligned} v_1(t) &= E_{MAX}(t) \\ &\stackrel{\text{def}}{=} \max\{E_{OX}(V_{GS}(\tau)), \tau \in [0,t]\} \end{aligned} \tag{6.40}$$

The equations to produce $v_1(t)$ during the circuit simulation are

$$v_1(t) = \max\{v_3(t), v_2(t)\} \tag{6.41a}$$

$$v_3(t) = \frac{1}{2}\left[v_4(t) + E_O + \delta E_O + \sqrt{(v_4(t) - E_O + \delta E_O)^2 + 4E_O\ \delta E_O}\right] \tag{6.41b}$$

$$v_4(t) = \frac{V_{GS}(t) + V_O}{A\ t_{OX}} \times 1 \times 10^{-8} \quad [\text{MV/cm}] \tag{6.41c}$$

To emphasize, the field quantities in MV/cm that physically exist in the gate dielectric are calculated and stored as voltage quantities in V in the simulation circuit. The "max" function in Equation 6.41a means selecting the larger value between the two variables v_3 (evaluated here) and v_2 ("present" voltage across C_{P1}). Equation 6.41b is the same smoothing function as Equation 6.37a. Equation 6.41c is the linear field model similar to Equation 6.37b but in MV/cm. The parameter values for V_O, A, E_O and δE_O in Table 6.5 are used. E_O and δE_O are converted to MV/cm, and $t_{OX} = 3\,\text{nm}$. The series and parallel resistors R_{S1} and R_{P1} help avoid matrix singularities in the SPICE simulation. Their associated RC time constants are entirely artificially introduced, not related to the physical process, and intentionally made irrelevant to the simulated functional circuit's time scales.

The oxide is determined as "broken" when $v_1(t) = E_{MAX}(t) \geq E_{BD}$, and the equivalent gate resistance R_G changes from the "good" value to the "broken" value. This transition needs to be continuous and differentiable for a transient simulation. Therefore, R_G is

calculated with a sigmoid smoothing function as follows[28].

$$\begin{aligned} R_G(t) &= R_{BD} + \frac{R_{GOOD}}{1+\exp\left(\frac{E_{MAX}(t)-E_{BD}}{\delta E_{BD}}\right)} \\ &= R_{BD} + \frac{R_{GOOD}}{2}\left[1-\tanh\left(\frac{E_{MAX}(t)-E_{BD}}{2\delta E_{BD}}\right)\right] \end{aligned} \tag{6.42}$$

where $E_{MAX}(t)$, E_{BD} and δE_{BD} are in MV/cm. $E_{BD} = 8.188\,\text{MV/cm}$ and $R_{BD} = 307.1\,\text{k}\Omega$ are the average BD field and post-BD resistance for device "8a", respectively, found in Section 5.4. $R_{GOOD} = 1\,\text{G}\Omega$ is an arbitrarily set value for a "good" or "pristine" oxide. $\delta E_{BD} = 0.1\,\text{MV/cm}$ is arbitrarily chosen, too, to ensure a smooth yet rapid transition.

To construct the *black-box* circuit model for practical applications, a dependent source ("B source" in ngspice) is used, and the final top-level circuit for R_G is a voltage-dependent current source $I_G(t)$ with

$$I_G(t) = \frac{V_{GS}(t)}{R_G(t)} \tag{6.43}$$

The equivalent circuit for the $R_G(V_{GS})$ and $E_{OX}(V_{GS})$ input-output relationships is simulated in ngspice with a DC voltage sweep. The results are shown in Figure 6.21.

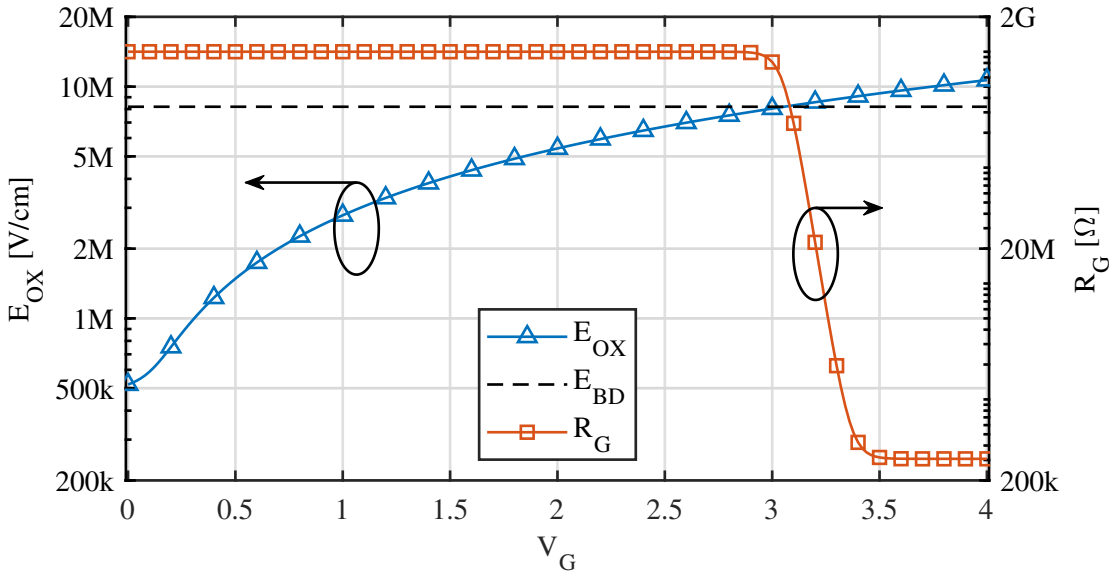


Figure 6.21: DC simulation results of the FinFET gate oxide RBD model.

In the FinFET, E_{OX} reaches $E_{BD} \approx 8\,\text{MV/cm}$ at around $V_{GS} = 3$–$3.5\,\text{V}$, when R_G

[28] Note that the quantity "R_{BD}" here means the post-BD gate resistance or "R_G" used in Section 5.4, while the abbreviation styled as "RBD" always stands for rapid oxide breakdown.

decreases as the oxide breaks down.

A test time-dependent simulation is demonstrated by simply applying a fluctuating $V_{GS}(t)$ and measuring $I_G(t)$. Only R_G is present in the circuit; the standard MOSFET (M1) is not included to exclude any transient gate current reported by the MOSFET capacitance models. The gate voltage V_{GS} (stated below) is designed to be non-negative (unipolar), rapidly fluctuating between a high value and zero, and slowly increasing its amplitude.

$$V_{GS}(t) = \left(\frac{t}{t_M}\right)\frac{A}{2}\left\{1+\sin\left[2\pi\left(\frac{3}{4}+f_c\,t\right)\right]\right\} \tag{6.44}$$

where $t_M = 5\,\text{ms}$, $A = 5\,\text{V}$, and $f_c = 10\,\text{kHz}$, so $V_{GS}(t)$ starts from $0\,\text{V}$ at $t = 0$, reaches its peak values every $f_c^{-1} = 0.1\,\text{s}$, and the peak value reaches $5\,\text{V}$ at about $t \approx 5\,\text{ms}$. The resulting waveforms are presented in Figure 6.22.

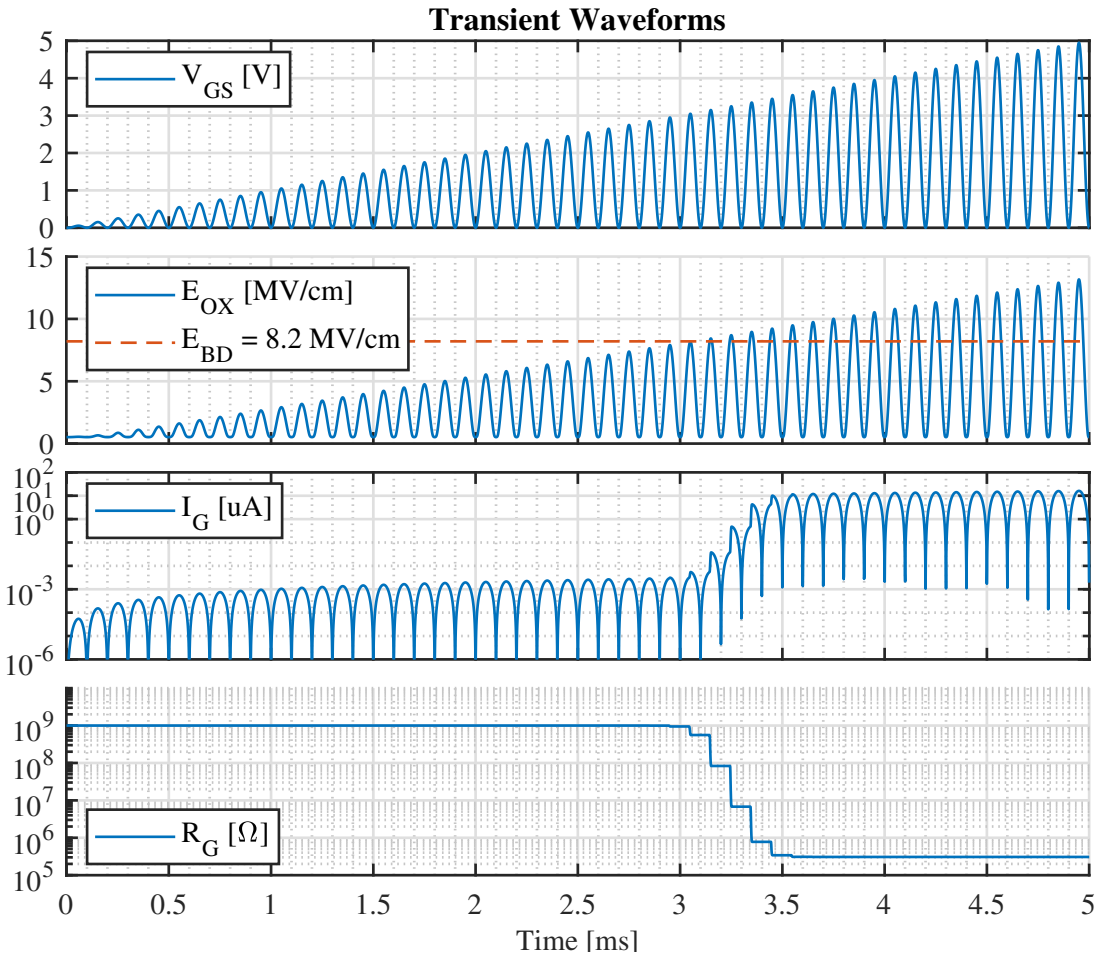


Figure 6.22: Transient simulation waveforms of the extracted oxide RBD model.

Prior to $t \approx 3\,\text{ms}$, the oxide field is below the BD threshold $E_{OX} \leq E_{BD}$, and the conductive current is low, $I_G\,(t \lesssim 3\,\text{ms}) \leq 2\,\text{nA}$. At $t \approx 3\,\text{ms}$ when E_{BD} has been reached, the oxide starts to breakdown and never goes back to the "good" condition. After $t \gtrsim 3.5\,\text{ms}$, the oxide is considered as "broken", and the conductive current $I_G\,(t \gtrsim 3.5\,\text{ms}) \geq 10\,\mu\text{A}$ is more than 5,000 times higher than before.

To summarize, a compact circuit model is proposed to assist the FinFET oxide RBD evaluation in transient simulations, based on our FinFET device-level simulations and our in-house stress-test experiments on planar devices. In addition to this model, one can improve the accuracy by including the dependency on the drain terminal voltage and experiments on ultra-thin dielectrics.

Chapter 7: Summary

This work has studied the vulnerabilities in MOSFET devices and circuits when transient terminal voltage and current disruptions are present when the system is exposed to electromagnetic interference. We analyze and evaluate the vulnerabilities in two categories, the non-permanent Soft Errors and the permanent Hard Failures.

The Soft Errors can temporarily disrupt the circuit's function by changing the signal-level behaviors, such as inducing bit errors and analog signal distortions, which can be unexpected by design. When MOSFET devices experience voltage-current surges, Soft Errors may occur and be intensified by the device's internal mechanisms, including the Snapback phenomenon. Due to the complications related to MOSFET vulnerabilities, the circuit may generate more errors or higher distortions than expected by regular MOSFET models. The circuit may even fall into an abnormal state, causing the system to freeze or malfunction until a hard reset. We explain the Soft Error vulnerability with physical mechanisms such as impact ionization and the parasitic bipolar junction structure, which are activated only under disrupted situations. Device-level simulations are performed to evaluate these mechanisms inside actual MOSFET devices.

The Hard Failures can permanently damage the devices and circuits and even prevent normal functions. Under EMI conditions, the terminal voltage-current disruptions are usually temporary but may be enough to induce Hard Failures. One particular kind of Hard Failures is studied, the gate oxide or dielectric breakdown. After reviewing the physical explanations from the literature, we establish a single criterion of rapid oxide breakdown, the oxide electric field.

The knowledge and computational simulations of Soft Errors and Hard Failures are verified by experiments. For the Soft Errors, we measure the terminal characteristics of a single off-the-shelf MOSFET device. Besides the regular I-V relation, the Snapback

phenomenon is observed under extremely high terminal voltages, which can occur when EMI-induced disruptions are present. For the Hard Failures, we stress test MOSFET devices fabricated on our own. The oxide breakdown results are analyzed statistically. To extend our knowledge of oxide breakdown to modern devices with thin oxides, we simulate a FinFET first at the device level and then including quantum-mechanical correction at the mesoscopic level. Typical device characteristics of the FinFET are also observed and discussed.

Circuit-level models are developed to represent the two types of vulnerabilities in MOSFET devices. These models are compatible with SPICE circuit simulations and can be used as an add-on to the regular MOSFET model. The data from both of our experiments are used to verify and calibrate the model parameters.

Example circuit simulations are performed. For the Soft Errors, we simulate basic practical circuits using the realistic model parameters extracted from our experiments. We demonstrate the Soft Errors under transient voltage-current disruptions, presumably induced by EMI. Bit errors and analog waveform distortions are observed. We compare the intensity of Soft Errors when the MOSFET's internal complications, collectively as the Snapback phenomenon, is and is not included in the simulation model. Due to Snapback, the vulnerable circuits experience additional power consumption and non-linear change in analog amplification gain. For the Hard Failures, we create a simple equivalent circuit representing the parallel (or leakage) resistance of the gate dielectric after it has broken down once the internal field has reached the threshold we have experimentally established.

The proposed compact circuit-level models are based on physical mechanisms and are consistent with detailed device-level simulations which are based on the distributed semiconductor device system of partial-differential equations. Experiment or simulation data can be used to extract model parameters for particular device designs of interest. We believe that the same phenomena related to MOSFET vulnerabilities can still happen and even intensify when the device dimensions scale down, as the internal electric field (one of

the causes to Soft Errors and Hard Failures) generally remains the same or increases. The proposed workflow can potentially help MOSFET device and circuit designers by bringing insights into the device's behaviors when EMI-induced temporary disruptions are present in the circuit.

References

[1] Toshio Sudo, Hideki Sasaki, Norio Masuda, and James L Drewniak. "Electromagnetic interference (EMI) of system-on-package (SOP)". In: *IEEE Transactions on Advanced Packaging* 27.2 (2004), pp. 304–314.

[2] Luc B Gravelle and Perry F Wilson. "EMI/EMC in printed circuit boards-a literature review". In: *IEEE Transactions on electromagnetic compatibility* 34.2 (1992), pp. 109–116.

[3] J-J Laurin, Safwat G Zaky, and Keith G Balmain. "On the prediction of digital circuit susceptibility to radiated EMI". In: *IEEE Transactions on Electromagnetic Compatibility* 37.4 (1995), pp. 528–535.

[4] Cemal Nalbantoglu, Thorsten Kiehl, Ralf God, Thiemo Stadtler, Robert Kebel, and Renke Bienert. "Electromagnetic compatibility (EMC) for integration and use of near field communication (NFC) in aircraft". In: *Aerospace EMC (Aerospace EMC), 2016 ESA Workshop on*. IEEE. 2016, pp. 1–6.

[5] Wilmar Heuvelman, Rick Janssen, and Ralph Prestros. "FEM modeling of gigahertz TEM cells for susceptibility analysis of RFID products". In: *Microwave Conference (EuMC), 2012 42nd European*. IEEE. 2012, pp. 530–533.

[6] Richard Redl. "Electromagnetic environmental impact of power electronics equipment". In: *Proceedings of the IEEE* 89.6 (2001), pp. 926–938.

[7] V Serrao, A Lidozzi, L Solero, and A Di Napoli. "EMI characterization and communication aspects for power electronics in hybrid vehicles". In: *Power Electronics and Applications, 2007 European Conference on*. IEEE. 2007, pp. 1–10.

[8] Hyok J Song, Joseph S Colburn, Hui P Hsu, and Richard W Wiese. "Modeling effect of lightning induced EMP on wire harness in automobiles". In: *IEEE Antennas and Propagation Society International Symposium*. Vol. 2. IEEE; 1999. 2005, p. 383.

[9] Eugenio Mattei, Elena Lucano, Federica Censi, Michele Triventi, and Giovanni Calcagnini. "Provocative testing for the assessment of the electromagnetic interference of RFID and NFC readers on implantable pacemaker". In: *IEEE Transactions on Electromagnetic Compatibility* 58.1 (2016), pp. 314–322.

[10] Meng Zhang, Anand Raghunathan, and Niraj K Jha. "Trustworthiness of Medical Devices and Body Area Networks." In: *Proceedings of the IEEE* 102.8 (2014), pp. 1174–1188.

[11] Chen Wang, Marco Leone, James L Drewniak, and Antonio Orlandi. “Coupling between differential signals and the DC power-bus in multilayer PCBs”. In: *IEEE transactions on advanced packaging* 28.2 (2005), pp. 337–345.

[12] Paul E Dodd, Marty R Shaneyfelt, James A Felix, and James R Schwank. “Production and propagation of single-event transients in high-speed digital logic ICs”. In: *IEEE Transactions on Nuclear Science* 51.6 (2004), pp. 3278–3284.

[13] Young Chung, Hongzhong Xu, Richard Ida, and Bob Baird. “Snapback breakdown dynamics and ESD susceptibility of LDMOS”. In: *2006 IEEE International Reliability Physics Symposium Proceedings*. IEEE. 2006, pp. 352–355.

[14] Salvatore Lombardo, James H Stathis, Barry P Linder, Kin Leong Pey, Felix Palumbo, and Chih Hang Tung. “Dielectric breakdown mechanisms in gate oxides”. In: *Journal of Applied Physics* 98.12 (2005), p. 12.

[15] Guido T Sasse, Fred G Kuper, and Jurriaan Schmitz. “MOSFET degradation under RF stress”. In: *IEEE transactions on electron devices* 55.11 (2008), pp. 3167–3174.

[16] Binhong Li, Nestor Berbel, Alexandre Boyer, Sonia Bendhia, and Raul Fernandez-Garcia. “Study of the impact of hot carrier injection to immunity of MOSFET to electromagnetic interferences”. In: *Microelectronics reliability* 51.9-11 (2011), pp. 1557–1560.

[17] Paul B Sparks, HARRY G MONO, Kenneth H Joyner, and Michael P Wood. “The safety of digital mobile cellular telephones with minute ventilation rate adaptive pacemakers”. In: *Pacing and clinical electrophysiology* 19.10 (1996), pp. 1451–1455.

[18] HY David Yang and Ronald Kollman. “Analysis of high-power RF interference on digital circuits”. In: *Electromagnetics* 26.1 (2006), pp. 87–102.

[19] Kyechong Kim and Agis A Iliadis. “Operational upsets and critical new bit errors in CMOS digital inverters due to high power pulsed electromagnetic interference”. In: *Solid-state electronics* 54.1 (2010), pp. 18–21.

[20] Zahra Abedi, Sameer Hemmady, Thomas Antonsen, Edl Schamiloglu, and Payman Zarkesh-Ha. “Electromagnetic Compatibility in Leakage Current of CMOS Integrated Circuits”. In: *2019 International Symposium on Electromagnetic Compatibility-EMC EUROPE*. IEEE. 2019, pp. 765–768.

[21] William Liu. *MOSFET Models for SPICE Simulation including BSIM3v3 and BSIM4*. Wiley-Interscience Publication, 2001.

[22] S. M. Sze and Kwok K. Ng. “Physics of Semiconductor Devices”. In: Third. John Wiley & Sons, Inc., 2007. Chap. 2.4.3.

[23] Binhong Li, Nestor Berbel, Alexandre Boyer, Sonia Bendhia, and Raul Fernandez-Garcia. “Study of the impact of hot carrier injection to immunity of MOSFET to electromagnetic interferences”. In: *Microelectronics reliability* 51.9-11 (2011), pp. 1557–1560.

[24] Young Chung, Hongzhong Xu, Richard Ida, and Bob Baird. “Snapback breakdown dynamics and ESD susceptibility of LDMOS”. In: *Reliability Physics Symposium Proceedings, 2006. 44th Annual., IEEE International*. IEEE. 2006, pp. 352–355.

[25] Fu-Chieh Hsu, Ping-Keung Ko, Simon Tam, Chenming Hu, and Richard S Muller. “An analytical breakdown model for short-channel MOSFET’s”. In: *IEEE Transactions on Electron Devices* 29.11 (1982), pp. 1735–1740.

[26] Mario Pinto-Guedes and Philip C. Chan. “A Circuit Simulation Model for Bipolar-Induced Breakdown in MOSFET”. In: *IEEE Transaction on Computer-Aided Design* 7.2 (Feb. 1988), pp. 289–294.

[27] Yuanzhong Zhou, Duane Connerney, Ronald Carroll, and Timwah Luk. “Modeling MOS snapback for circuit-level ESD simulation using BSIM3 and VBIC models”. In: *Quality of Electronic Design, 2005. ISQED 2005. Sixth International Symposium on*. IEEE. 2005, pp. 476–481.

[28] Marty Johnson, Roger Cline, Scott Ward, and Joe Schichl. *White paper: Latch-Up*. Tech. rep. SCAA124. Texas Instruments, Apr. 2015. URL: http://www.ti.com/lit/wp/scaa124/scaa124.pdf.

[29] Ajith Amerasekera, Sridhar Ramaswamy, Mi-Chang Chang, and Charvaka Duvvury. “Modeling MOS snapback and parasitic bipolar action for circuit-level ESD and high current simulations”. In: *Reliability Physics Symposium, 1996. 34th Annual Proceedings., IEEE International*. IEEE. 1996, pp. 318–326.

[30] William Shockley. “Problems related to pn junctions in silicon”. In: *Solid-State Electronics* 2.1 (1961), 35IN961–60IN1067.

[31] Chang-Hoon Choi, Kwang-Hoon Oh, Jung-Suk Goo, Zhiping Yu, and Robert W Dutton. “Direct tunneling current model for circuit simulation”. In: *International Electron Devices Meeting 1999. Technical Digest (Cat. No. 99CH36318)*. IEEE. 1999, pp. 735–738.

[32] A. Gehring, H. Kosina, T. Grasser, and S. Selberherr. “Consistent Comparison of Tunneling Models for Device Simulation”. In: 4th European Workshop on Ultimate Integration of Silicon. Udine, Italy, 2003, pp. 131–134.

[33] Andreas Gehring, Tibor Grasser, B-H Cheong, and Siegfried Selberherr. “Design optimization of multi-barrier tunneling devices using the transfer-matrix method”. In: *Solid-State Electronics* 46.10 (2002), pp. 1545–1551.

[34] Takashi Ando, Ninad D Sathaye, Kota VRM Murali, and Eduard A Cartier. “On the Electron and Hole Tunneling in a HfO2 Gate Stack With Extreme Interfacial-Layer Scaling”. In: *IEEE electron device letters* 32.7 (2011), pp. 865–867.

[35] Chung-Kuang Huang and Neil Goldsman. “Non-equilibrium modeling of tunneling gate currents in nanoscale MOSFETs”. In: *Solid-State Electronics* 47.4 (2003), pp. 713–720.

[36] Sivakumar Mudanai, Yang-Yu Fan, Qiqing Ouyang, Al F Tasch, and Sanjay Kumar Banerjee. “Modeling of direct tunneling current through gate dielectric stacks”. In: *IEEE Transactions on electron devices* 47.10 (2000), pp. 1851–1857.

[37] Wei-Kai Shih, Everett X Wang, Srinivas Jallepalli, Francisco Leon, Christine M Maziar, and Al F Tasch Jr. “Modeling gate leakage current in nMOS structures due to tunneling through an ultra-thin oxide”. In: *Solid-State Electronics* 42.6 (1998), pp. 997–1006.

[38] Masataka Hirose. “Electron tunneling through ultrathin SiO2”. In: *Materials Science and Engineering: B* 41.1 (1996), pp. 35–38.

[39] N Goldsman and Jeffrey Frey. “Efficient and accurate use of the energy transport method in device simulation”. In: *IEEE Transactions on Electron Devices* 35.9 (1988), pp. 1524–1529.

[40] José P Rino, Ingvar Ebbsjö, Rajiv K Kalia, Aiichiro Nakano, and Priya Vashishta. “Structure of rings in vitreous SiO 2”. In: *Physical Review B* 47.6 (1993), p. 3053.

[41] Al-Moatasem El-Sayed, Matthew B Watkins, Alexander L Shluger, and Valeri V Afanas’ev. “Identification of intrinsic electron trapping sites in bulk amorphous silica from ab initio calculations”. In: *Microelectronic engineering* 109 (2013), pp. 68–71.

[42] A Padovani, DZ Gao, AL Shluger, and L Larcher. “A microscopic mechanism of dielectric breakdown in SiO2 films: An insight from multi-scale modeling”. In: *Journal of Applied physics* 121.15 (2017), p. 155101.

[43] A Kerber, A Vayshenker, D Lipp, T Nigam, and E Cartier. “Impact of charge trapping on the voltage acceleration of TDDB in metal gate/high-k n-channel MOSFETs”. In: *2010 IEEE International Reliability Physics Symposium*. IEEE. 2010, pp. 369–372.

[44] David Z Gao, Al-Moatasem El-Sayed, and Alexander L Shluger. “A mechanism for Frenkel defect creation in amorphous SiO2 facilitated by electron injection”. In: *Nanotechnology* 27.50 (2016), p. 505207.

[45] JW McPherson and HC Mogul. “Underlying physics of the thermochemical E model in describing low-field time-dependent dielectric breakdown in SiO 2 thin films”. In: *Journal of Applied Physics* 84.3 (1998), pp. 1513–1523.

[46] Yuan Taur and Tak H Ning. “Fundamentals of modern VLSI devices”. In: Cambridge university press, 1998. Chap. 4.1.1.

[47] Il-Kwon Oh, Jukka Tanskanen, Hanearl Jung, Kangsik Kim, Mi Jin Lee, Zonghoon Lee, Seoung-Ki Lee, Jong-Hyun Ahn, Chang Wan Lee, Kwanpyo Kim, et al. “Nucleation and growth of the HfO2 dielectric layer for graphene-based devices”. In: *Chemistry of Materials* 27.17 (2015), pp. 5868–5877.

[48] Luca Vandelli, Andrea Padovani, Luca Larcher, and Gennadi Bersuker. “Microscopic modeling of electrical stress-induced breakdown in poly-crystalline hafnium

oxide dielectrics". In: *IEEE transactions on electron devices* 60.5 (2013), pp. 1754–1762.

[49] Peter M Zeitzoff. "MOSFET scaling trends and challenges through the end of the roadmap". In: *Proceedings of the IEEE 2004 Custom Integrated Circuits Conference (IEEE Cat. No. 04CH37571)*. IEEE. 2004, pp. 233–240.

[50] Ali Khakifirooz and Dimitri A Antoniadis. "MOSFET performance scaling—Part II: Future directions". In: *IEEE Transactions on Electron Devices* 55.6 (2008), pp. 1401–1408.

[51] *International Technology Roadmap for Semiconductors 2.0 2015 Edition: Executive Report*. Whitepaper. International Roadmap Committee, 2018.

[52] *International Roadmap for Devices and Systems 2017 Edition: More Moore*. Whitepaper. International Roadmap Committee, 2018.

[53] *International Roadmap for Devices and Systems 2020 Update: More Moore*. Whitepaper. International Roadmap Committee, 2020.

[54] Jean-Pierre Colinge et al. *FinFETs and other multi-gate transistors*. Vol. 73. Springer, 2008.

[55] A VY Thean, ZH Shi, L Mathew, T Stephens, H Desjardin, C Parker, T White, M Stoker, L Prabhu, R Garcia, et al. "Performance and variability comparisons between multi-gate FETs and planar SOI transistors". In: *Electron Devices Meeting, 2006. IEDM'06. International*. IEEE. 2006, pp. 1–4.

[56] Felix Bloch and prepared by John D. Walecka. "Fundamentals of Statistical Mechanics". In: Stanford University Press, 1989. Chap. V.13.

[57] Carlo Jacoboni and Lino Reggiani. "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials". In: *Reviews of modern Physics* 55.3 (1983), p. 645.

[58] Neil Goldsman, Yu-Jen Wu, and Jeffrey Frey. "Efficient calculation of ionization coefficients in silicon from the energy distribution function". In: *Journal of Applied Physics* 68.3 (1990), pp. 1075–1081.

[59] N Goldsman, L Henrickson, and J Frey. "Reconciliation of a hot-electron distribution function with the lucky electron-exponential model in silicon". In: *IEEE Electron Device Letters* 11.10 (1990), pp. 472–474.

[60] K Taniguchi, M Yamaji, K Sonoda, T Kunikiyo, and C Hamaguchi. "Monte Carlo study of impact ionization phenomena in small geometry MOSFET's". In: *Electron Devices Meeting, 1994. IEDM'94. Technical Digest., International*. IEEE. 1994, pp. 355–358.

[61] E Cartier, MV Fischetti, EA Eklund, and FR McFeely. "Impact ionization in silicon". In: *Applied Physics Letters* 62.25 (1993), pp. 3339–3341.

[62] Paolo Nenzi and Holger Vogt. *Ngspice Users Manual Version 23*. 2011.

[63] Eric W Weisstein. *Projection*. Feb. 4, 2021. URL: https://mathworld.wolfram.com/Projection.html (visited on 02/04/2021).

[64] Yuan Taur and Tak H Ning. "Fundamentals of modern VLSI devices". In: Cambridge university press, 1998. Chap. 2.4.1.

[65] F Nouketcha, A Lelis, R Green, Y Cui, C Darmody, and N Goldsman. "Detailed Study of Breakdown Voltage and Critical Field in Wide Bandgap Semiconductors". In: *2019 IEEE 7th Workshop on Wide Bandgap Power Devices and Applications (WiPDA)*. IEEE. 2019, pp. 200–207.

[66] John L Moll. *Physics of semiconductors*. McGraw-Hill, 1964, pp. 216–219.

[67] F. L. L. Nouketcha, Y. Cui, A. Lelis, R. Green, C. Darmody, J. Schuster, and N. Goldsman. "Investigation of Wide- and Ultrawide-Bandgap Semiconductors From Impact-Ionization Coefficients". In: *IEEE Transactions on Electron Devices* 67.10 (2020), pp. 3999–4005.

[68] David A. Gates. "Design-Oriented Mixed-Level Circuit and Device Simulation". PhD thesis. EECS Department, University of California, Berkeley, 1993. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/1993/2382.html.

[69] A Akturk, S Potbhare, J Booz, N Goldsman, D Gundlacha, R Nandwanab, and K Mayaramb. "CoolSPICE: SPICE for Extreme Temperature Range Integrated Circuit Design and Modeling". In: *Proceedings of Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD)*. 2012, pp. 63–66.

[70] S. M. Sze. "Physics of Semiconductor Devices". In: Second. John Wiley & Sons, Inc., 1981. Chap. 3.2.2.

[71] Yuan Taur and Tak H Ning. "Fundamentals of modern VLSI devices". In: Cambridge university press, 1998. Chap. 6.3.3.

[72] Andrei Vladimirescu and Sally Liu. *The Simulation of MOS Integrated Circuits Using SPICE2*. Technical Report. EECS Department, University of California, Berkeley, Feb. 1980. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/1980/ERL-m-80-7.pdf (visited on 08/17/2020).

[73] Jerry Mar, Sheau-Suey Li, and Swei-Yam Yu. "Substrate current modeling for circuit simulation". In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 1.4 (1982), pp. 183–186.

[74] Yuan Taur and Tak H Ning. "Fundamentals of modern VLSI devices". In: Cambridge university press, 1998. Chap. 3.1.2.

[75] XB Chen, ZQ Song, and ZJ Li. "Optimization of the drift region of power MOSFET's with lateral structures and deep junctions". In: *IEEE Transactions on Electron Devices* 34.11 (1987), pp. 2344–2350.

[76] Enrico Sangiorgi. "Latch-up in CMOS circuits: A review". In: *Transactions on Emerging Telecommunications Technologies* 1.3 (1990), pp. 337–349.

[77] Digi-Key. *CD4007UBE*. https://www.digikey.com/product-detail/en/texas-instruments/CD4007UBE/296-3501-5-ND/376600. Online; accessed 28-February-2020.

[78] Adelmo Ortiz-Conde, FJ Garcıa Sánchez, Juin J Liou, Antonio Cerdeira, Magali Estrada, and Y Yue. "A review of recent MOSFET threshold voltage extraction methods". In: *Microelectronics Reliability* 42.4-5 (2002), pp. 583–596.

[79] R Van Overstraeten and H De Man. "Measurement of the ionization rates in diffused silicon pn junctions". In: *Solid-State Electronics* 13.5 (1970), pp. 583–608.

[80] C. Darmody and N. Goldsman. "Incomplete ionization in aluminum-doped 4H-silicon carbide". In: *Journal of Applied Physics* 126.14 (2019), p. 145701. DOI: 10.1063/1.5120707. eprint: https://doi.org/10.1063/1.5120707. URL: https://doi.org/10.1063/1.5120707.

[81] Gabriel Vasilescu. "Electronic Noise and Interfering Signals". In: Springer-Verlag Berlin Heidelberg, 2005. Chap. 11.

[82] Ernest Y Wu and Jordi Suñé. "Generalized hydrogen release-reaction model for the breakdown of modern gate dielectrics". In: *Journal of Applied Physics* 114.1 (2013), p. 014103.

[83] JW McPherson, RB Khamankar, and A Shanware. "Complementary model for intrinsic time-dependent dielectric breakdown in SiO 2 dielectrics". In: *Journal of Applied Physics* 88.9 (2000), pp. 5351–5359.

[84] Yee-Chia Yeo, Qiang Lu, and Chenming Hu. "MOSFET gate oxide reliability: Anode hole injection model and its applications". In: *International journal of high speed electronics and systems* 11.03 (2001), pp. 849–886.

[85] Yuan Taur and Tak H Ning. "Fundamentals of modern VLSI devices". In: Cambridge university press, 1998. Chap. 2.3.1.

[86] Wenping Wang, Vijay Reddy, Anand T Krishnan, Rakesh Vattikonda, Srikanth Krishnan, and Yu Cao. "Compact modeling and simulation of circuit reliability for 65-nm CMOS technology". In: *IEEE Transactions on Device and Materials Reliability* 7.4 (2007), pp. 509–517.

[87] DL Griscom and M Cook. "29Si superhyperfine interactions of the E' center: a potential probe of range-II order in silica glass". In: *Journal of non-crystalline solids* 182.1-2 (1995), pp. 119–134.

[88] Massimo V Fischetti, Donelli J DiMaria, SD Brorson, TN Theis, and JR Kirtley. "Theory of high-field electron transport in silicon dioxide". In: *Physical Review B* 31.12 (1985), p. 8124.

[89] S-H Lo, DA Buchanan, Y Taur, and W Wang. "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET's". In: *IEEE Electron Device Letters* 18.5 (1997), pp. 209–211.

[90] Arvind Kumar, Sandip Mondal, and KSR Koteswara Rao. “Structural, electrical, band alignment and charge trapping analysis of nitrogen-annealed Pt/HfO 2/p-Si (100) MIS devices”. In: *Applied Physics A* 122.12 (2016), pp. 1–8.

[91] G Bersuker, N Chowdhury, C Young, D Heh, D Misra, and R Choi. “Progressive breakdown characteristics of high-k/metal gate stacks”. In: *Reliability physics symposium, 2007. proceedings. 45th annual. ieee international*. IEEE. 2007, pp. 49–54.

[92] Eduard Cartier and Andreas Kerber. “Stress-induced leakage current and defect generation in nFETs with HfO 2/TiN gate stacks during positive-bias temperature stress”. In: *Reliability Physics Symposium, 2009 IEEE International*. IEEE. 2009, pp. 486–492.

[93] R Pagano, Salvatore Lombardo, Felix Palumbo, Paul Kirsch, SA Krishnan, C Young, Rino Choi, Gennadi Bersuker, and James H Stathis. “A novel approach to characterization of progressive breakdown in high-k/metal gate stacks”. In: *Microelectronics Reliability* 48.11-12 (2008), pp. 1759–1764.

[94] Jun-Young Park, Dong-Il Moon, Myeong-Lok Seol, Choong-Ki Kim, Chang-Hoon Jeon, Hagyoul Bae, Tewook Bang, and Yang-Kyu Choi. “Self-curable gate-all-around MOSFETs using electrical annealing to repair degradation induced from hot-carrier injection”. In: *IEEE Transactions on Electron Devices* 63.3 (2016), pp. 910–915.

[95] KIKUO Yamabe and KENJI Taniguchi. “Time-dependent-dielectric breakdown of thin thermally grown SiO 2 films”. In: *IEEE Transactions on Electron Devices* 32.2 (1985), pp. 423–428.

[96] A Cattaneo, S Pinarello, J-E Mueller, and Robert Weigel. “MOSFET degradation under DC and RF Fowler-Nordheim stress”. In: *2014 44th European Solid State Device Research Conference (ESSDERC)*. IEEE. 2014, pp. 230–233.

[97] Minitab, Inc. *Minitab 18 Statistical Software*. Version 18.1. 2017. URL: www.minitab.com.

[98] Yuan Taur and Tak H Ning. “Fundamentals of modern VLSI devices”. In: Cambridge university press, 1998. Chap. 3.2.1.

[99] LD Yau. “A simple theory to predict the threshold voltage of short-channel IGFET’s”. In: *Solid-State Electronics* 17.10 (1974), pp. 1059–1063.

[100] B Sampath Kumar, Milova Paul, Harald Gossner, and Mayank Shrivastava. “Physical Insights into the ESD behavior of Drain Extended FinFETs”. In: *2018 40th Electrical Overstress/Electrostatic Discharge Symposium (EOS/ESD)*. IEEE. 2018, pp. 1–7.

[101] Seshadri Kolluri, Kazuhiko Endo, Eiichi Suzuki, and Kaustav Banerjee. “Modeling and analysis of self-heating in FinFET devices for improved circuit and EOS/ESD performance”. In: *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*. IEEE. 2007, pp. 177–180.

[102] S Lee, R Wachnik, P Hyde, L Wagner, J Johnson, A Chou, A Kumar, S Narasimha, T Standaert, B Greene, et al. "Experimental analysis and modeling of self heating effect in dielectric isolated planar and fin devices". In: *VLSI Technology (VLSIT), 2013 Symposium on*. IEEE. 2013, T248–T249.

[103] Eun-Ae Chung, Kab-Jin Nam, Toshiro Nakanishi, Sungil Park, Hongseon Yang, Thomas Kauerauf, Guangfan Jiao, Dong-won Kim, Ki Hyun Hwang, Hyejin Kim, et al. "Investigation of hot carrier degradation in bulk FinFET". In: *Reliability Physics Symposium (IRPS), 2017 IEEE International*. IEEE. 2017, XT–6.

[104] Jian-Hsing Lee, Manjunatha Prabhu, Konstantin Korablev, Jagar Singh, Mahadeva Iyer Natarajan, and Shesh Mani Pandey. "Methodology to achieve planar technology-like ESD performance in FINFET process". In: *Reliability Physics Symposium (IRPS), 2015 IEEE International*. IEEE. 2015, 3F–3.

[105] Wen-Shiang Liao, Yie-Gie Liaw, Mao-Chyuan Tang, Sandipan Chakraborty, and Chee Wee Liu. "Investigation of reliability characteristics in NMOS and PMOS FinFETs". In: *IEEE Electron Device Letters* 29.7 (2008), pp. 788–790.

[106] S Ramey, A Ashutosh, C Auth, J Clifford, M Hattendorf, J Hicks, R James, A Rahman, V Sharma, A St Amour, et al. "Intrinsic transistor reliability improvements from 22nm tri-gate technology". In: *Reliability Physics Symposium (IRPS), 2013 IEEE International*. IEEE. 2013, pp. 4C–5.

[107] B Sampath Kumar, Milova Paul, Harald Gossner, and Mayank Shrivastava. "Physical insights into the ESD behavior of drain extended FinFETs (DeFinFETs) and unique current filament dynamics". In: *IEEE Transactions on Electron Devices* 67.7 (2020), pp. 2717–2724.

[108] Tai-su Park, Euijoon Yoon, and Jong-Ho Lee. "A 40 nm body-tied FinFET (OMEGA MOSFET) using bulk Si wafer". In: *Physica E: Low-dimensional Systems and Nanostructures* 19.1-2 (2003), pp. 6–12.

[109] Chang Yong Kang, Changwoo Sohn, Rock-Hyun Baek, Chris Hobbs, Paul Kirsch, and Raj Jammy. "Effects of layout and process parameters on device/circuit performance and variability for 10nm node FinFET technology". In: *VLSI Technology (VLSIT), 2013 Symposium on*. IEEE. 2013, T90–T91.

[110] T Kanemura, T Izumida, N Aoki, M Kondo, S Ito, T Enda, K Okano, H Kawasaki, A Yagishita, A Kaneko, et al. "Improvement of drive current in bulk-FinFET using full 3D process/device simulations". In: *Simulation of Semiconductor Processes and Devices, 2006 International Conference on*. IEEE. 2006, pp. 131–134.

[111] Xue Shao and Zhiping Yu. "Nanoscale FinFET simulation: A quasi-3D quantum mechanical model using NEGF". In: *Solid-State Electronics* 49.8 (2005), pp. 1435–1445.

[112] Zhihao Yu, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. "Effects of fin shape on sub-10 nm FinFETs". In: *Journal of Computational Electronics* 14.2 (2015), pp. 515–523.

[113] Eric W Weisstein. *Airy Functions*. MathWorld–A Wolfram Web Resource. Oct. 23, 2020. URL: https://mathworld.wolfram.com/AiryFunctions.html (visited on 11/17/2020).

[114] Chien-Wei Lee and Jenn-Gwo Hwu. "Quantum-mechanical calculation of carrier distribution in MOS accumulation and strong inversion layers". In: *AIP Advances* 3.10 (2013), p. 102123. DOI: 10.1063/1.4826886. eprint: https://doi.org/10.1063/1.4826886. URL: https://doi.org/10.1063/1.4826886.

[115] Eric W Weisstein. *Fermi-Dirac Distribution*. MathWorld–A Wolfram Web Resource. Oct. 15, 2020. URL: https://mathworld.wolfram.com/Fermi-DiracDistribution.html (visited on 10/15/2020).

[116] Christopher Darmody. "DFT and Related Modeling of Post-Silicon Valence 4 Materials: SiC and Ge". PhD thesis. University of Maryland (College Park, Md.), 2020. Chap. 4.6. URL: https://doi.org/10.13016/cy6v-ukj4.