MODELING SELF-HEATING AND NEGATIVE BIAS TEMPERATURE INSTABILITY IN FINFETS

by

David E. Sommer

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Applied Physics).

Golden, Colorado

Date _____

Signed: _____
David E. Sommer

Signed: _____
Dr. Lincoln D. Carr
Thesis Advisor

Signed: _____
Dr. Carole Graas
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____
Dr. Tom Furtak
Professor and Head
Department of Physics

ABSTRACT

Negative bias temperature instability (NBTI) is a significant wearout mechanism responsible for the degradation of crucial device performance characteristics in p-doped metal-oxide-semiconductor field-effect transistor(pMOSFET)technologies. As MOSFET dimensions are pushed deeper into the limits of scaling, NBTI is expected to become an even greater reliability concern due to higher electric fields and greater amounts of self-heating. For new device architectures such as the FinFET, predictive modeling of NBTI is both essential to the long-term success of the technology and is still in its nascent stages of development. This work constitutes a first step toward establishing a simulation infrastructure for the predictive modeling of NBTI wearout in the FinFET.

In the following, we consider the effects of NBTI degradation and self-heating on the threshold voltage characteristics for p-doped silicon FinFET structures using technology computer-aided design (TCAD) simulation. We employ a two-stage hole trapping model for NBTI wearout and solve for a coupled set of moment equations derived from the Boltzmann transport equation on a finite element mesh. For both two- and three-dimensional FinFET structures, we observe a non-monotonic variation in the threshold voltage degradation rate not observed in typical NBTI stress experiments on planar MOSFET structures, and we find that monotonicity is restored for high temperatures and asymmetric stress configurations. We further conclude that the self-heating of the lattice is significant only for asymmetric stress configurations. Finally, we hypothesize that overestimates in the charge carrier velocity and underestimates in the lattice self-heating contribute significantly to the observed non-monotonic behavior, and this provides impetus and directions for future work.

TABLE OF CONTENTS

ABSTRACT iii

LIST OF FIGURES vi

LIST OF TABLES xiii

LIST OF ABBREVIATIONS xiv

ACKNOWLEDGMENTS xv

DEDICATION xvi

CHAPTER 1 INTRODUCTION 1

1.1 Moore's Law and the Need for FinFETs 3

1.2 An Overview of Device Reliability 5

1.3 The Evolving Importance of Self-Heating 10

CHAPTER 2 THEORY AND BACKGROUND 13

2.1 Nanoscale Transport 14

2.1.1 Microscopic Foundations: The Quantum Boltzmann Equation . 15

2.1.2 Semiclassical Transport: The Boltzmann Transport Equation . . 17

2.1.3 Device Equations: Moments of the Boltzmann Transport Equation 24

2.2 Phonon Scattering and Self-heating 28

2.2.1 Microscopic Foundations 28

2.2.2 Monte Carlo Simulation and Nonequilibrium Phonon Distributions 31

2.3 Negative Bias Temperature Instability and the Two-stage Model 33

2.3.1 Microscopic Foundations: Defects and Nonradiative Capture Processes . . . 36

2.3.2 Two-stage Negative Bias Temperature Instability: The E'/P_b Switching Trap Model . . . 44

2.3.3 The Reaction-Diffusion Framework and Unified Reliability Models . . . 47

2.4 The Finite-Element Simulation Environment and Negative Bias Temperature Instability Stressing . . . 48

2.4.1 Simulating Negative Bias Temperature Instability . . . 50

CHAPTER 3 SIMULATIONS AND RESULTS . . . 53

3.1 Case 1: 2D pFinFET . . . 53

3.1.1 Symmetric Stress Configuration . . . 55

3.1.2 Asymmetric Stress Configuration . . . 83

3.1.3 The Effect of Thermal Boundary Conditions . . . 89

3.1.4 Simulation Convergence and Numerical Error . . . 93

3.2 Case 2: 2D pMOSFET . . . 108

3.3 Case 3: 3D pFinFET . . . 109

CHAPTER 4 CONCLUSIONS . . . 129

4.1 Research Results and Outlook . . . 129

4.2 Recommendations for Future Work . . . 132

REFERENCES CITED . . . 134

LIST OF FIGURES

Figure 1.1 Generic Tri-Gate SOI FinFET 2

Figure 1.2 The Bathtub Curve 7

Figure 2.1 Charge transport regimes in solids: classical → semiclassical → quantum mechanical 15

Figure 2.2 Phonon dispersion curves for silicon 31

Figure 2.3 Nonradiative MPE capture process coordinate diagram 38

Figure 2.4 Two-stage NBTI model 44

Figure 2.5 Example finite element mesh for 2D FinFET structure 49

Figure 2.6 Heuristic NBTI simulation steps 51

Figure 3.1 2D FinFET structure 54

Figure 3.2 Temperature variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $V_{\mathrm{g}} = -0.5$ V 55

Figure 3.3 Temperature variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $V_{\mathrm{g}} = -1.0$ V 56

Figure 3.4 Temperature variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $V_{\mathrm{g}} = -1.5$ V 56

Figure 3.5 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 325$ K 57

Figure 3.6 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 350$ K 58

Figure 3.7 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 375$ K 58

Figure 3.8 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 400$ K 59

Figure 3.9 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 450$ K 59

Figure 3.10 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 500$ K 60

Figure 3.11 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 550$ K 60

Figure 3.12 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 600$ K 61

Figure 3.13 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 650$ K 61

Figure 3.14 Variation of ΔV_{th} after 1 s symmetric stress for varying gate voltage and temperature 62

Figure 3.15 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during asymmetric stress ($V_{\mathrm{g}} = V_d$), $T_{\mathrm{ext}} = 350$ K 62

Figure 3.16 Recovery of V_{th} degradation following symmetric stress, $T_{\mathrm{ext}} = 325$ K 63

Figure 3.17 Recovery of V_{th} degradation following symmetric stress, $T_{\mathrm{ext}} = 350$ K 64

Figure 3.18 Recovery of V_{th} degradation following symmetric stress, $T_{\mathrm{ext}} = 375$ K 64

Figure 3.19 Recovery of V_{th} degradation following symmetric stress, $T_{\mathrm{ext}} = 400$ K 65

Figure 3.20 Temperature variation of $\langle f_1 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -0.5$ V 66

Figure 3.21 Temperature variation of $\langle f_1 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -1.0$ V 66

Figure 3.22 Temperature variation of $\langle f_1 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -1.5$ V 67

Figure 3.23 Temperature variation of $\langle f_2 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -0.5$ V 67

Figure 3.24 Temperature variation of $\langle f_2 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -1.0$ V 68

Figure 3.25 Temperature variation of $\langle f_2 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -1.5$ V 68

Figure 3.26 Temperature variation of $\langle f_3 \rangle$ during symmetric stress, $V_\mathrm{g} = -0.5$ V 69

Figure 3.27 Temperature variation of $\langle f_3 \rangle$ during symmetric stress, $V_\mathrm{g} = -1.0$ V 69

Figure 3.28 Temperature variation of $\langle f_3 \rangle$ during symmetric stress, $V_\mathrm{g} = -1.5$ V 70

Figure 3.29 Temperature variation of $\langle f_4 \rangle$ during symmetric stress, $V_\mathrm{g} = -0.5$ V 70

Figure 3.30 Temperature variation of $\langle f_4 \rangle$ during symmetric stress, $V_\mathrm{g} = -1.0$ V 71

Figure 3.31 Temperature variation of $\langle f_4 \rangle$ during symmetric stress, $V_\mathrm{g} = -1.5$ V 71

Figure 3.32 Temperature variation of Q during symmetric stress, $V_\mathrm{g} = -0.5$ V . 73

Figure 3.33 Temperature variation of Q during symmetric stress, $V_\mathrm{g} = -1.0$ V . 73

Figure 3.34 Temperature variation of Q during symmetric stress, $V_\mathrm{g} = -1.5$ V . 74

Figure 3.35 Gate voltage variation of $\langle f_1 \rangle$ during symmetric stress, $T_\mathrm{ext} = 350$ K 75

Figure 3.36 Gate voltage variation of $\langle f_1 \rangle$ during symmetric stress, $T_\mathrm{ext} = 550$ K 75

Figure 3.37 Gate voltage variation of $\langle f_2 \rangle$ during symmetric stress, $T_\mathrm{ext} = 350$ K 76

Figure 3.38 Gate voltage variation of $\langle f_2 \rangle$ during symmetric stress, $T_\mathrm{ext} = 550$ K 76

Figure 3.39 Gate voltage variation of $\langle f_3 \rangle$ during symmetric stress, $T_\mathrm{ext} = 350$ K 77

Figure 3.40 Gate voltage variation of $\langle f_3 \rangle$ during symmetric stress, $T_\mathrm{ext} = 550$ K 77

Figure 3.41 Gate voltage variation of $\langle f_4 \rangle$ during symmetric stress, $T_\mathrm{ext} = 350$ K 78

Figure 3.42 Gate voltage variation of $\langle f_4 \rangle$ during symmetric stress, $T_\mathrm{ext} = 550$ K 78

Figure 3.43 Gate voltage variation of Q during symmetric stress, $T_\mathrm{ext} = 350$ K . 79

Figure 3.44 Gate voltage variation of Q during symmetric stress, $T_\mathrm{ext} = 550$ K . 79

Figure 3.45 Hole velocity $|\boldsymbol{v}_\mathrm{h}|$ under symmetric stress 80

Figure 3.46 Electric field $|\boldsymbol{E}|$ under symmetric stress 81

Figure 3.47 Hole density n_h under symmetric stress 81

Figure 3.48 Hole energy ϵ_h under symmetric stress 82

Figure 3.49 Hole current density $|\boldsymbol{J}_{\mathrm{h}}|$ under symmetric stress 83

Figure 3.50 Lattice temperature T_{L} under symmetric stress 84

Figure 3.51 Gate voltage variation of $\langle f_1 \rangle$ during asymmetric stress, $T_{\mathrm{ext}} = 350$ K 85

Figure 3.52 Gate voltage variation of $\langle f_2 \rangle$ during asymmetric stress, $T_{\mathrm{ext}} = 350$ K 86

Figure 3.53 Gate voltage variation of $\langle f_3 \rangle$ during asymmetric stress, $T_{\mathrm{ext}} = 350$ K 87

Figure 3.54 Gate voltage variation of $\langle f_4 \rangle$ during asymmetric stress, $T_{\mathrm{ext}} = 350$ K 88

Figure 3.55 Gate voltage variation of Q during asymmetric stress, $T_{\mathrm{ext}} = 350$ K 89

Figure 3.56 Hole velocity $|\boldsymbol{v}_{\mathrm{h}}|$ under asymmetric stress 90

Figure 3.57 Electric field $|\boldsymbol{E}|$ under asymmetric stress 91

Figure 3.58 Hole density n_{h} under asymmetric stress 92

Figure 3.59 Hole energy ϵ_{h} under asymmetric stress 93

Figure 3.60 Hole current density $|\boldsymbol{J}_{\mathrm{h}}|$ under asymmetric stress 94

Figure 3.61 Lattice temeperature T_{L} under asymmetric stress 95

Figure 3.62 Hole velocity $|\boldsymbol{v}_{\mathrm{h}}|$: symmetric versus asymmetric stress with Neumann boundary conditions 96

Figure 3.63 Hole density n_{h}: symmetric versus asymmetric stress with Neumann boundary conditions 97

Figure 3.64 Hole energy ϵ_{h}: symmetric verses asymmetric stress with Neumann boundary conditions 98

Figure 3.65 Hole current density n_{h} under asymmetric stress ($V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K) with Neumann boundary conditions 98

Figure 3.66 Lattice temperature T_{L}: symmetric versus asymmetric stress with Neumann conditions 99

Figure 3.67 Relative $\Delta V_{\mathrm{th}}(t)$ for SHE and Hydrodynamic simulations ($V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 325$ K) 100

Figure 3.68 Hole velocity $|\boldsymbol{v}_{\mathrm{h}}|$: symmetric versus asymmetric stress with Hydrodynamic model 101

Figure 3.69 Hole density n_{h}: symmetric versus asymmetric stress with Hydrodynamic model 102

Figure 3.70 Hole temperature T_{h}: symmetric versus asymmetric stress with Hydrodynamic model 103

Figure 3.71 Hole current density $|\boldsymbol{J}_{\mathrm{h}}|$: symmetric versus asymmetric stress with Hydrodynamic model 104

Figure 3.72 Lattice temperature T_{L}: symmetric versus asymmetric stress with Hydrodynamic model 105

Figure 3.73 Failed versus successful convergence during quasistationary ramp for 2D FinFET NBTI stress simulations ($V_{\mathrm{g}} = -2.0$ V) 106

Figure 3.74 Finite element mesh and doping concentration for 2D MOSFET structure 108

Figure 3.75 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 300$ K 109

Figure 3.76 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 350$ K 110

Figure 3.77 Gate voltage variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $T_{\mathrm{ext}} = 375$ K 110

Figure 3.78 Gate voltage variation of $\langle f_1 \rangle$ during symmetric stress, $T_{\mathrm{ext}} = 375$ K 111

Figure 3.79 Gate voltage variation of $\langle f_2 \rangle$ during symmetric stress, $T_{\mathrm{ext}} = 375$ K 111

Figure 3.80 Gate voltage variation of $\langle f_3 \rangle$ during symmetric stress, $T_{\mathrm{ext}} = 375$ K 112

Figure 3.81 Gate voltage variation of $\langle f_4 \rangle$ during symmetric stress, $T_{\mathrm{ext}} = 375$ K 112

Figure 3.82 Gate voltage variation of Q during symmetric stress, $T_{\text{ext}} = 375$ K 113

Figure 3.83 2D Planar MOSFET under symmetric stress ($V_{\text{g}} = -1.5$ V, $T_{\text{ext}} = 375$ K) 114

Figure 3.84 3D FinFET structure with finite element mesh 115

Figure 3.85 Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 325$ K 116

Figure 3.86 Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 375$ K 116

Figure 3.87 Gate voltage variation of $\Delta V_{\text{th}}(t)$ during asymmetric stress ($V_{\text{g}} = V_d$), $T_{\text{ext}} = 375$ K 117

Figure 3.88 Gate voltage variation of $\langle f_1 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K 117

Figure 3.89 Gate voltage variation of $\langle f_2 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K 118

Figure 3.90 Gate voltage variation of $\langle f_3 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K 118

Figure 3.91 Gate voltage variation of $\langle f_4 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K 119

Figure 3.92 Gate voltage variation of Q during symmetric stress, $T_{\text{ext}} = 375$ K 119

Figure 3.93 Gate voltage variation of $\langle f_1 \rangle$ during asymmetric stress, $T_{\text{ext}} = 375$ K 120

Figure 3.94 Gate voltage variation of $\langle f_2 \rangle$ during asymmetric stress, $T_{\text{ext}} = 375$ K 121

Figure 3.95 Gate voltage variation of $\langle f_3 \rangle$ during asymmetric stress, $T_{\text{ext}} = 375$ K 121

Figure 3.96 Gate voltage variation of $\langle f_4 \rangle$ during asymmetric stress, $T_{\text{ext}} = 375$ K 122

Figure 3.97 Gate voltage variation of Q during asymmetric stress, $T_{\text{ext}} = 375$ K 122

Figure 3.98 Hole velocity $|\boldsymbol{v}_{\text{h}}|$ under symmetric stress ($T_{\text{ext}} = 375$ K) 123

Figure 3.99 Electric field $|\boldsymbol{E}|$ under symmetric stress ($T_{\mathrm{ext}} = 375$ K) 124

Figure 3.100 Hole density n_{h} under symmetric stress ($T_{\mathrm{ext}} = 375$ K) 124

Figure 3.101 Hole temperature T_{h} under symmetric stress ($T_{\mathrm{ext}} = 375$ K) . . . 125

Figure 3.102 Hole current density $|\boldsymbol{J}_{\mathrm{h}}|$ under symmetric stress ($T_{\mathrm{ext}} = 375$ K) . 126

Figure 3.103 Lattice temperature T_{L} under symmetric stress ($T_{\mathrm{ext}} = 375$ K) . . 126

LIST OF TABLES

Table 2.1 Comparison between hydrodynamic and SHE models 52

Table 3.1 Simulation cases 53

LIST OF ABBREVIATIONS

Back-end-of-line . . . . . BEOL

Boltzmann transport equation . . . . . BTE

Electromigration . . . . . EMG

Front-end-of-line . . . . . FEOL

Hot carrier injection . . . . . HCI

Metal-oxide-semiconductor field-effect transistor . . . . . MOSFET

Multiphonon emission . . . . . MPE

Multiphonon-field-assisted tunneling . . . . . MPFAT

Negative bias temperature instability . . . . . NBTI

Quantum Boltzmann equation . . . . . QBE

Short-channel effect . . . . . SCE

Spherical harmonic expansion . . . . . SHE

Stress-induced voiding . . . . . SV

Technology computer-aided design . . . . . TCAD

Time-dependent dielectric breakdown . . . . . TDDB

ACKNOWLEDGMENTS

If I were to catalogue the conversations, ideas, and inspirations that contributed to this work and to the development of my academic career, the length of such a document would rival the length of this thesis. Suffice it to say that any thanks I could give here is inevitably inadequate, but I hope that this forgivable, as there is still much that lies ahead.

First, I would like to thank the physics department at the Colorado School of Mines for their kindness and intelligence as I waded into the world of physics. Foremost, I would like to thank my advisor Lincoln Carr for not only challenging and inspiring me with intriguing problems, but for helping me develop the scientific abilities to address them. In the same vein, I would like to thank Scott Strong for his patience and insightful conversations as I began my first foray into research. Furthermore, I would like to thank the members of my committee, Reuben Collins, Tim Ohno, and David Wood, for the time they spent on my thesis as well as for the brief and interesting conversations and classes I had with them.

This work would not have been possible without the vision and guidance of my advisor Carole Graas and the support of IBM. To her I owe the deepest gratitude for trusting me to attempt the daunting task of breaking new ground in the reliability physics community. Additionally, I would like to thank Giuseppe LaRosa for challenging me and motivating this work. Furthermore, I am grateful for the support and friendliness of Theodoros Anemikos, Ernest Wu, Steve Furkay, Jeff Johnson, Andres Bryant, Ron Bolam, and Connie Dolinger who helped make my experience at IBM incredibly valuable. Finally, I would like to thank Brendon Murphy of Synopsys, whose tireless efforts to help me develop and understand the TCAD simulation made this work possible.

Maybe a twelvemonth since
Suddenly I began,
In scorn of this audience,
Imagining a man
And his sun-freckled face,
And grey Connemara cloth,
Climbing up to a place
Where stone is dark under froth,
And the down turn of his wrist
When the flies drop in the stream:
A man who does not exist,
A man who is but a dream;
And cried, 'Before I am old
I shall have written him one
Poem maybe as cold
And passionate as the dawn.'

- William Butler Yeats
From "The Fisherman"

CHAPTER 1
INTRODUCTION

As conventional planar metal-oxide-semiconductor field-effect transistor (MOSFET) devices reach their functional limit for dimensional scaling, the semiconductor industry must employ novel materials and device architectures to continue its trend toward smaller and faster transistors. With the introduction of new transistor architectures comes the need for a new developmental methodology. The existence of this need can be understood in two ways. First, as the technology delves further into the nanometer regime, the emergence of quantum mechanical phenomena becomes an increasingly relevant concern. Second, if the technology is to be considered viable and manufacturable in the long-term, then it must demonstrate a certain degree of reliability, and the reliability metrics of these novels architectures are still unclear, if not under-emphasized. This thesis addresses the second issue. Given that the evolution of the MOSFET up to this point has been based on a standard planar geometry, issues of performance and reliability analysis could be reduced to understanding how device behavior conformed to scaling. However, with new potential architectures emerging, meeting performance standards is not enough; a proper understanding of reliability must also occur in the nascent stages of device development.

The FinFET structure (Figure 1.1) shows particular viability as a new transistor model for gate lengths below 20 nm, admitting both relatively easy process integration and greater control of short-channel effects (SCE). However, the unconventional geometry of the device raises new questions about process control and calibration, in addition to performance and reliability. With respect to the latter, one must account for all three dimensions of the conduction channel and the greater surface area of the dielectric interface in order to properly model the behavior of the device. Further-

more, the shorter length scales of the channel raise new questions about electrical and thermal transport in confined geometries. A particular consequence of the FinFET structure is the emergence of localized hot spots near the drain region of the device [9]. The localization of heating has important implications not only for carrier mobility in the channel (and thus device performance) but also for the activation of intrinsic wearout mechanisms. A precise understanding of the effects of self-heating on FinFET wearout, under both stress and normal operating conditions, could dictate design parameters to be implemented in the early stages of development, which would enhance the lifetime of the device. This work constitutes a first step toward characterizing the nature of self-heating and its effects on a particular wearout mechanism, namely the negative bias temperature instability (NBTI), in FinFET structures.

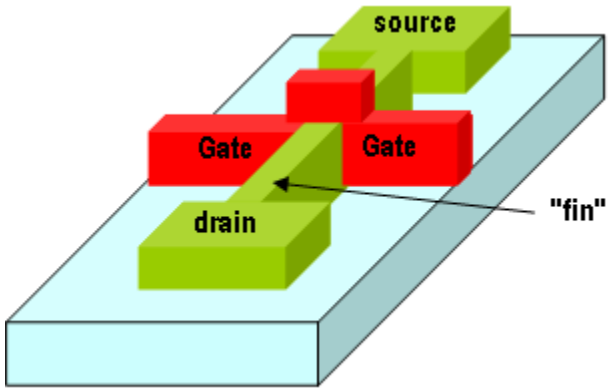


Figure 1.1: Generic Tri-Gate SOI FinFET

In the following, we will review several perspectives on charge transport, self-heating and NBTI wearout in nanostructures, and then demonstrate how the confluence of these previously distinct ideas lends itself to a new way of thinking about device reliability. In particular, we will discuss the results that have been gathered so far, their implications, and then what still remains to be done.

1.1 Moore's Law and the Need for FinFETs

In 1965, Gordon E. Moore described a trend encapsulating the continuous improvement of integrated circuit performance as reflected by the progressive scaling of the transistor and the increase of on-chip component density. Simply stated, the prediction set down by Moore maintains that the number of devices on a chip will quadruple while doubling transistor performance every three years. Miraculously, this prediction, now commonly referred to as Moore's law, has provided an accurate roadmap for transistor development and the semiconductor industry. While the demand for faster and better performing technology has continued to provide much of the impetus for this innovation, such progress would not be possible without the ability to circumvent some of the limitations of fundamental physics encountered on ever smaller length scales [1]. Indeed, until recently, the industry has relied on the shrinking of the conventional planar MOSFET as the cornerstone of this innovative drive. Quantum-mechanical tunneling through thin gate oxide layers and from source to drain (the *punch-through* effect) as well as the complexity involved in controlling dopant levels throughout the device are just some of the challenges encountered at each new technology node. While the essential structure of the planar MOSFET has remained the same, the employment of additional structures and materials, such as high-k metal gates, silicon-on-insulator (SOI) design and channel straining, have allowed the MOSFET to be scaled to smaller than 50 nm. The present smallest transistor size is 22 nm [1].

Nonetheless, the extent to which the planar MOSFET can be scaled will reach its practical and theoretical limit at a gate length of approximately 20 nm due to the gate's inability to control the channel region of the device. At such length scales, the electric field between the source and drain regions gives rise to SCE. A new transistor architecture is required to circumvent these problems. Since the 1980's, it has been suggested that a a dual-gate FET, in which two opposite sides of the channel are

bounded by a gate, is a possible way to extend transistor scaling while still gaining in performance [1, 5]. By increasing the relative surface area of the gate-channel interface, the channel of the device is better controlled by the gate potential. This tighter coupling between gate and channel allows for greater control of SCE with less reliance on doping, as well as steeper subthreshold slope for better switching performance. Lowered doping has the additional benefit of improving carrier transport through the channel, as the probability for Coulomb scattering has been reduced, as well as lowering the local electric field near the drain.

Despite the clear potential advantages of the dual-gate FET, controlling the alignment of the two gates has been a significant roadblock to its realization. A fruitful way to circumvent this problem is to construct a thin vertical channel on a bulk or SOI substrate and then wrap the gate around the exposed sides of the channel [5]. In this way, one can have either a tri-gate structure or, if a hard mask is added to the top part of the channel, a dual-gate structure. This self-aligned transistor structure is aptly called the FinFET, 'Fin' referring to the thin, vertical orientation of the channel (Figure 1.1). Nonetheless, significant challenges remain. Maintaining uniform channel thickness while reducing line-edge roughness, controlling the Fin pitch, properly aligning source and drain regions, optimizing the height of the fin, and determining the size and placement of contacts and interconnects continue to be on-going problems for processing. In terms of performance, properly understanding carrier mobility and scattering, the effects of doping and straining, and the nature of parasitic resistances and capacitances that arise from the three-dimensional geometry of the device (i.e., effects of the corners) are also in their developing stages [2]. Questions of reliability are also crucial to the success of the FinFET as the next step in transistor scaling but have, as yet, received less attention.

1.2 An Overview of Device Reliability

The success of a novel technology not only demands that it meets performance and processing criteria, but also that it will function properly for a reasonable amount of time. In order to develop reliability criteria and to predict the useful lifetime of a device, one must be able to characterize the various defects and wearout mechanisms that contribute to its failure. Conventionally, for the planar MOSFET, this is done empirically by stress testing a large sample of devices and fitting their failure data to a statistical model [6]. It has often been found that for newer technologies, the results of this stress testing require retroactive changes to processing parameters and steps, which ultimately complicates manufacturing and reduces yield. Thus in the early stages of development for novel structures such as the FinFET, it becomes particularly crucial to be able to predict device reliability behavior concurrently with performance and process feasibility in order to ensure the device's long-term viability.

The use of device simulation concurrent with device development and integration is now common practice in the semiconductor industry [3]. To match the demands of scaling reflected in Moore's law, simulation has become an indispensable component for saving development time as well as resources and cost, allowing for the optimization of device parameters and processes. Indeed, the increasing use of simulation to reduce costs, improve efficiency, and predict the emergence of new phenomena is not unique to the semiconductor industry. For example, computer-aided-design software has long been used to streamline and standardize production in construction and mechanical manufacturing. IBM's development of deep computing to understand complex social and biological systems is yet another example of the growing use of simulation. Likewise, the implementation of "computation thinking" to facilitate engineering innovation and scientific discovery is the basis for the National Science Foundation cyber-enabled discovery and innovation program [4]. These examples are representative of a growing trend of using computational resources to understand

increasingly complex systems with multiple degrees of freedom.

The simulation and prediction of device reliability requires an understanding of the physics of semiconductor defects and wearout. Several wearout mechanisms, including NBTI, depend on the interaction of charge carriers with defects in the semiconductor structure. Thus, we briefly review the various defect structures that can occur in the material lattice [34, 35]:

1. **Point Defects:** Localized, point-like disruptions in the discrete lattice symmetry.
    - **Vacancy:** the absence of an atom in the lattice.
    - **Interstitial:** the presence of an additional atom in the lattice (occupying an 'interstitial' site).
    - **Substitutional:** the replacement of a regular host atom A by an atom B.
        - **Antisite:** a substitutional defect in which a regular host atom A is replaced by another host atom B.
    - **Shallow Impurity:** a hydrogen-like point defect characterized by a highly delocalized wave function extending over many primitive cells.
    - **Deep Center:** point defects characterized by highly localized wave functions and significant lattice relaxation. Their energy levels often occur near the middle of the bandgap.
2. **Line Defects:** Extended, line-like disruption in the discrete lattice symmetry.
    - **Dislocation:** the displacement of one or more rows of atoms in the lattice.
    - **Grain Boundary:** a localized lattice region rotated in orientation with respect to the surrounding crystalline structure.
3. **Complexes:** small clusters of point defects.

- **Frenkel defect pair:** a vacancy-interstitial complex formed by the displacement of a lattice atom to a nearby interstitial site.

The instantaneous failure rate, or hazard function, over time for a given set of devices follows the distribution given in Figure 1.2, referred to as the bathtub curve. Early failures (i.e., infant mortalities) occur primarily due to significant manufacturing defects, such as the improper deposition or bonding of materials, the deposition of excess material causing an errant conduction pathway, or the introduction of foreign particles during processing. Once these devices have failed, the failure rate levels off for the useful lifetime of the device. At the end of the distribution, we find that wearout mechanisms begin to increase the failure rate until 100 percent of the set has failed. When wearout begins to play a significant role and which mechanisms dominate the wearout process dictate the stress and normal use conditions for the device.

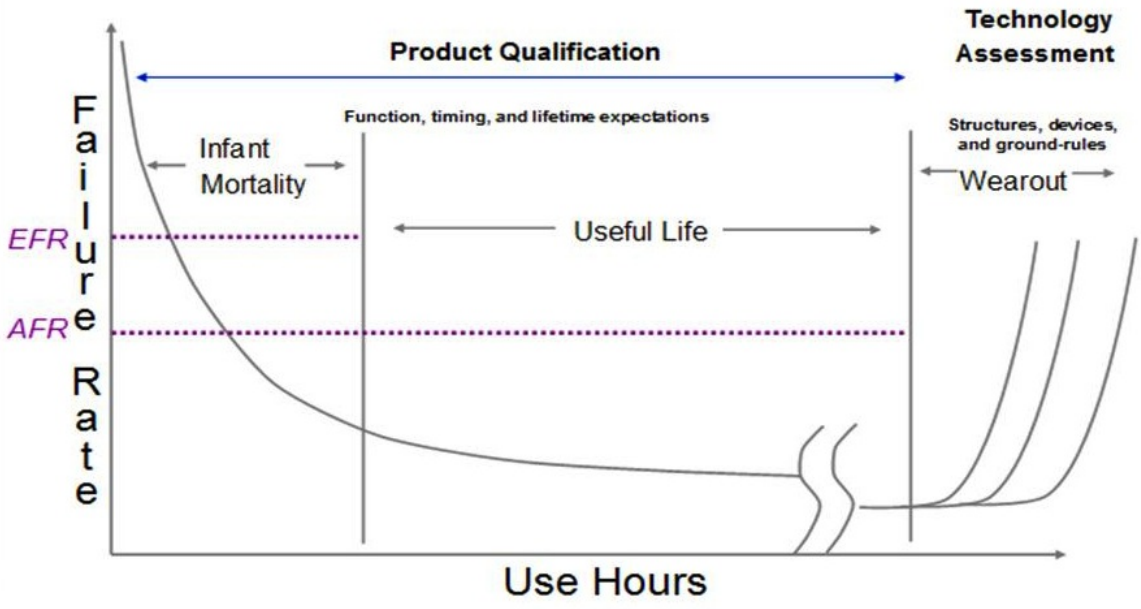


Figure 1.2: The Bathtub Curve

There are several intrinsic wearout mechanisms that contribute to the eventual failure of a complementary metal-oxide-semiconductor (CMOS) transistor device, and we describe some of the more significant ones below [6]:

- **Negative Bias Temperature Instability** (NBTI) refers to the degradation of threshold voltage and current-voltage (I-V) characteristics of a device due to the build-up of positive charges at the gate dielectric-channel interface and the generation of positively charged interface states. This mechanism is thermally activated and nonlinearly dependent on the electric field in the gate oxide.

- **Hot Carrier Injection** (HCI) refers to the degradation caused by the scattering of high-energy charge carriers (i.e., charge carriers whose effective temperature is higher than the lattice temperature, called *hot carriers*) at device interfaces and out of the channel. HCI typically occurs under high electric fields or from the impact of high energy radiation, such as alpha particles from cosmic ray showers [7]. The generation of hot carriers can affect the saturation condition of the device, as well as generate dangling bonds at interfaces (interface states) and substrate current due to impact ionization.

- **Time-Dependent Dielectric Breakdown** (TDDB) refers to the formation of a conduction pathway through the gate dielectric caused by the tunneling of charge carriers from the channel into the dielectric. This effect is particularly sensitive to oxide thickness and surface area contact with the channel, as well as the dielectric constant of the gate oxide material, the gate bias, and the temperature of the device.

- **Electromigration** (EMG) refers to the diffusion of metal particles through back-end-of-line (BEOL) metal interconnects and across material boundaries due the presence of an electric field and the apparent fluidity of metal particles assisted by defects, such as dislocations and grain boundaries.

- **Stress-Induced Voiding** (SV) refers to the formation of voids in BEOL metal contacts and structures due to mismatches in stress conditions and thermal coefficients at material interfaces and grain boundaries.

We can give a general classification to the various wearout mechanisms in semiconductor CMOS technology according to the reproducible nature of the degradation under stress [6]. *Parametric mechanisms* are those mechanisms that induce the same degree of degradation for identical structures under identical stress configurations. Thus, any variability in stress testing is the result of the intrinsic variability of the device structure. Empirical characterization of parametric mechanisms thus only requires small sample sizes given the degree of control over device processing. HCI, NBTI and electromigration are examples of parametric mechanisms. By contrast, *structural mechanisms* are those mechanisms for which device failure only depends on a certain region or structure within the device. Empirical characterization of structural mechanisms requires large sample sizes to determine that the degradation is structurally systematic. *Statistical mechanisms*, on the other hand, are classified by random behavior. Thus, identical structures will not fail in the same way for a statistical mechanism, but rather according to a random distribution (often a Weibull distribution [7]), and one is required to use large sample sizes to characterize the nature of the failure mode. TDDB is an example of a statistical mechanism. The underlying physics of a particular wearout mechanism determines its deterministic or statistical nature, and an understanding of the physics behind wearout is crucial to mitigating its effects. This underscores the need for predictive device reliability modeling.

Under stress conditions, understanding how the lattice heating of the device affects thermally dependent wearout mechanisms becomes crucial to predicting device reliability. This is especially so for novel structures like the FinFET, where the effects of the device geometry on self-heating may not be properly understood. Our

goal, then, is to provide a qualitative analysis of a particular wearout mechanism in a FinFET structure with self-heating. In the following, we will examine NBTI degradation, for which a direct physical connection to the self-heating mechanism can be established semiclassically.

1.3 The Evolving Importance of Self-Heating

For transistor devices deeply scaled beyond the 20 nm regime, the discrete nature of the atomic structure of the material plays an increasingly important role. When considering the transport of charge carriers through the device, one must not only be concerned with the wave nature of the charge carriers and the entailed quantum mechanical effects, but also with the quantization of lattice vibrations (phonons). The emission of phonons as electrons and holes scatter off the lattice dictates several important properties of device operation, including self-heating and switching time (determined by the saturation velocity) [38]. Conversely, the presence of phonons can limit electron and hole mobility, as well as activate new dynamic phenomena. Indeed, the interaction between electrons and phonons is so important that it is key component of the Bardeen-Cooper-Schrieffer (BCS) theory for superconductivity [32]. Nonetheless, for our purposes, we will be concerned with how carrier-phonon interactions affect the NBTI wearout of the device.

When a system is driven out of equilibrium, scattering processes act as pathways for momentum and energy relaxation, allowing the system to return to an equilibrium state. For example, phonon emission is often activated during charge transport when the average kinetic energy of the charges exceeds the average thermal energy of the lattice [34]. One can observe the macroscopic consequences of phonon emission when considering a steady state current in a metal in the presence of an external electric field. This steady state is maintained when the energy gained by the conduction electrons accelerated in the field is transferred to the lattice [8]. There are several different mechanisms for carrier-phonon scattering, which depend on the band structure of the

material and the phonon mode participating in the interaction. In general, a three dimensional lattice such as GaAs and Si can carry acoustic and optical phonon modes, from which one can distinguish transverse and longitudinal lattice motion.[1] Thus, we will distinguish between four modes: longitudinal acoustic (LA), transverse acoustic (TA), longitudinal optical (LO), and transverse optical (TO). Each mode involves a different momentum and energy transition when interacting with electrons.

We can alternatively characterize phonon scattering by comparing initial and final states. Intravalley scattering processes involve a momentum and/or energy transition within the same energy level or band. This process can be induced by both acoustic and optical phonons, as well as by impurity scattering. Intervalley scattering, on the other hand, involves a momentum and/or energy transition between energy levels or bands. This process can also be induced by both acoustical and optical phonons [30, 8]. For intervalley scattering, we can make the further distinction between normal and Umklapp processes. During normal processes, the electron momentum state remains in the first Brillouin zone, whereas during Umklapp processes the electron momentum state is scattered out of the first Brillouin zone [32].

For both polar (e.g. GaAs and InP) and covalent (e.g. Si) semiconductor materials at high temperatures, the dominant pathway for charge carrier energy loss is via the emission of longitudinal optical (LO) phonons, which then decay into acoustic phonons through the Klemens' channel [8]. Due to the non-zero group velocity of acoustic phonons, as compared to the near-zero group velocity of optical phonons, acoustic phonons act as the primary means of transferring heat away from the scattering center. However, due to a mismatch in the timescales for the transfer of carrier energy to LO phonons (on the order of 0.1 ps) and the decay of LO phonons into acoustic phonons (on the order of 10 ps), a state of thermal nonequilibrium can de-

[1]Here transverse and longitudinal are used in the typical sense for wave motion. For transverse waves, the wave displacement is perpendicular to the direction of propagation, while for longitudinal waves, the displacement is parallel to the direction of propagation.

velop between optical and acoustic phonon modes under strong electric fields. (A strong electric field is required to supply sufficient energy for the charge carriers to activate optical phonon modes.) This "phonon bottleneck" problem may have significant implications for charge carrier mobility under high-field transport, as well as the behavior of thermally dependent wearout mechanisms [9, 10, 11].

Few studies on FinFET reliability have yet been published. Scholten, *et al*, have experimentally characterized the effects of self-heating in SOI FinFETS on drain and drain-gate capacitances, extracting the frequency dependent thermal impedance of the device [18]. Using pulsed I-V measurements and S-parameter measurements, they found a significant rise in the device temperature ($\approx$ 80 K) compared to bulk planar MOSFETs, as well as an increase in thermal impedance at lower signal frequencies. Choi, *et al*, have investigated both hot carrier effects and NBTI in multi-gate CMOS FinFETs [19]. They observed a much higher sensitivity of NMOS FinFETs to HCI, while PMOS FinFETs experienced more significant degradation due to NBTI. Furthermore, they found that body-tied FinFET structures exhibit less NBTI degradation than SOI FinFET structures due to the presence of a grounded substrate. More importantly, NBTI degradation is increased for narrower fins, becoming the most significant limiting factor for device lifetime. Finally, Wang, *et al*, have tested the effects of self-heating due to ballistic phonons (i.e., the phonon bottleneck problem) in 90 nm planar pMOSFETs [20]. They found a higher than expect rise in channel temperature consistent with the generation of ballistic phonons and observed that NBTI degradation was significantly enhanced due to the presence of a localized hot spot near the drain.

CHAPTER 2

THEORY AND BACKGROUND

The predictive characterization of a reliability phenomena must rely on a sound, underlying physical model. Below, we discuss the Boltzmann transport equation (BTE) and its quantum mechanical foundations, which are essential for understanding and modeling the transport of charge carriers in two- and three-dimensional device geometries. We then discuss approximations (moments of the BTE) that can, and often must be made, in order to simulate device operation. Additionally, we provide a brief overview of the microscopic nature of self-heating, as well as of the complexity of its treatment in device simulation. Next, we discuss NBTI wearout in general and outline a two-stage charge-trapping model founded on the theory of charge capture in deep level defects by nonradiative multiphonon capture processes. We then make the case that these models lend themselves to a unified way of understanding NBTI and self-heating in FinFET structures. Finally, we discuss the finite element simulation environment, TCAD (technology computer-aided design), and sketch its implementation for solving the NBTI/self-heating problem for a FinFET structure.

It is important to note throughout this discussion that a certain number of tradeoffs between the computational efficiency of a simulation and the accuracy of the employed physical models must be made if one is to address such questions in a reasonable amount of time. Indeed, many sacrifices of model accuracy were required for a converged and reasonably efficient numerical solution. The purpose of this chapter, then, is to trace a logical path through the hierarchy of approximations that must be made in order to simulate the FinFET device. It is our hope that this work provides impetus for the continual improvement of simulating these phenomenon.

2.1 Nanoscale Transport

Our discussion of FinFET NBTI and self-heating begins with the theory of electron/hole transport, which will provide the foundation for the set of device equations characterizing the FinFET's operation. Indeed, the kinetics of charge carriers and the dissipative scattering processes that occur in the system dictate crucial device parameters such as conductivity and current-voltage relationships. Figure 2.1 shows a schematic representation of the different scaled perspectives with which we can treat charge transport phenomena. On the left, we have the classical picture, in which the solid is treated as a smooth background characterized by some macroscopic material parameters and transport occurs via a drift-diffusion process [38]. The next level in the hierarchy gives the semiclassical picture, wherein we treat the discrete nature of the lattice in an effective way and allow our models to account for more realistic interactions with the lattice. This regime is modeled almost exclusively by the Boltzmann transport equation (BTE), which describes the kinetics of a charge distribution function subject to classical Newtonian fields and effective scattering processes. This treatment becomes necessary in order to properly model aggressively scaled devices subject to high electric fields, and it assumes that the phase of the charge carriers is sufficiently randomized by numerous scattering events. For devices scaled to the order of the lattice spacing of the material, it becomes necessary to account for the wave nature of the charge carriers as well as of the lattice vibrations. In this quantum mechanical regime, represented in the right picture, one must be concerned with quantum phenomena such as the emergence of discrete energy subbands, phonon and charge confinement effects, and phase coherent transport. Such quantum mechanical approaches to charge transport include non-equilibrium Green's function (NEGF) methods (alternatively, the Kadanoff-Baym formalism), Wigner distribution functions, and the Kondo method [64, 32]. To facilitate our discussion of transport phenomena, we will only consider the transport and scattering of electrons

in semiconductor materials. The transport of holes can be treated in an analogous manner.

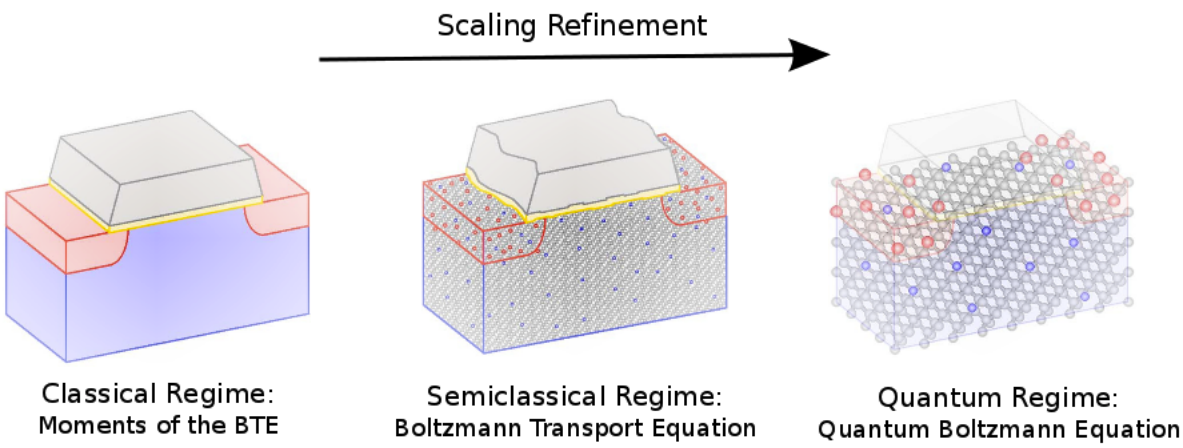


Figure 2.1: Charge transport regimes in solids: classical → semiclassical → quantum mechanical

2.1.1 Microscopic Foundations: The Quantum Boltzmann Equation

For completeness and to put the subsequent derivations on a firm theoretical foundation, we begin with a brief discussion of the quantum Boltzmann equation (QBE) and show how it is related to the semi-classical Boltzmann transport equation. The QBE describes the quantum transport of charges in solids under electromagnetic fields and readily includes low-order many-body correlations. It is generally derived using the nonequilibrium Green's function (NEGF) formalism, also referred to as the Kadanoff-Baym formalism, which allows one to directly account for quantum and nonequilibrium effects due to electromagnetic and phonon field interactions. Mahan provides two detailed derivations of the QBE for electrons in solids, the first including interactions with an electric field and the second including interactions with both an electric and magnetic field [22]. We briefly consider the former case.

The QBE is an equation of motion for the single-particle Green's function $G^{<}$. This Green's function is defined within second quantization as the correlation of a

field operator for a single fermion at two different positions and times

$$G^{<}(r_1, t_1; r_2, t_2) = i\langle \psi^{\dagger}(r_2, t_2)\psi(r_1, t_1)\rangle \ . \tag{2.1}$$

In order to obtain a quantum distribution function (the Wigner distribution function) similar to the distribution function satisfying the semiclassical Boltzmann transport equation,

$$\frac{\partial f}{\partial t} + v \cdot \nabla_r f + F \cdot \frac{1}{m}\nabla_v f + \left(\frac{\partial f}{\partial t}\right)_s = 0 \ , \tag{2.2}$$

we must make a change of coordinates. Thus, we combine both spatial and temporal components r_i and t_i into one spacetime coordinate $x_i = (r_i, t_i)$ and then define an effective center-of-mass coordinate system

$$(r, t) = x_1 - x_2 \ , \quad (R, T) = \frac{1}{2}(x_1 + x_2) \ . \tag{2.3}$$

We denote the new coordinate dependence of the Green's function as $G^{<}(r, t; R, T)$. Then, a Fourier transform in (r, t) yields

$$G^{<}(k, \omega; R, T) = \int d^3r e^{-ik\cdot r} \int dt e^{i\omega t} G^{<}(r, t; R, T) \ , \tag{2.4}$$

which can be related to the Wigner distribution function by

$$f(k, \omega; R, T) = -iG^{<}(k, \omega; R, T) \ . \tag{2.5}$$

The distribution $f(k, \omega; R, T)$ satisfies a form of the QBE very similar in structure to the semiclassical Boltzmann transport equation (2.2) [22]

$$\frac{\partial f}{\partial t} + v \cdot \nabla_r f + F \cdot \left[\frac{1}{m}\nabla_v + v\frac{\partial}{\partial \omega}\right] f + \left(\frac{\partial f}{\partial t}\right)_s = 0 \ . \tag{2.6}$$

Once the Wigner distribution function is known, or equivalently the Green's function $G^{<}$, we can compute all single-particle quantities of interest, such as the particle density, the particle current density, and the energy density [64]. This is an important feature of the QBE as well as the semiclassical Boltzmann formalism. An integral over the frequency recovers (2.2) and the semiclassical distribution function

$$f(r,v,t) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} f(mv,\omega;r,t) \ . \tag{2.7}$$

Indeed, the coarse-grained averaging used to obtain the semiclassical distribution removes the effects of the Pauli exclusion principle accounted for in the Wigner distribution function. While similar in structure, there are some important differences to note between the Wigner distribution function (2.5) and the semiclassical Boltzmann distribution in (2.2). First, in the quantum mechanical picture, there is no relationship between the energy E and the magnitude of the wave vector k, given that plane waves are not necessarily energy eigenstates. In the semiclassical picture, this distinction disappears, and the energy is defined implicitly based on the wave vector k. Second, because the Wigner function is not positive definite, it does not have a simple physical interpretation, whereas the semiclassical distribution is simply the number of particles at given spacetime coordinate with a given momentum [64].

2.1.2 Semiclassical Transport: The Boltzmann Transport Equation

The BTE is a kinetic equation that describes the time evolution of a single-particle distribution function f in six-dimensional phase space (i.e., three-dimensional physical space and three-dimensional momentum space) in response to the application of an external perturbation. Solutions to the BTE for a given system contain information about particle density and current, as well as energy and momentum transfer during charge transport, and thus f is a fundamental quantity for device simulation.[2] For transport in semiconductors, the Lorentz force (i.e., the existence of external electric and magnetic fields) typically acts as the external field driving perturbations of the system. However, in the following we will consider the case where only an electric field is present.

[2] f can be said to play an analogous role to the wave function ψ in quantum theory.

The validity of the BTE, in general, relies on the existence of two well-separated time scales characterizing, respectively, the duration of each collision τ_0 and the mean time between collisions τ. The latter is related to the mean free path of the particle, while the former depends on the coupling strength of the interaction constituting the collision. We thus take collisions to be essentially local and instantaneous, with $\tau_0 \ll \tau$. In practice, this requirement holds for metals and semiconductors [33, 37].

We can justify the semiclassical treatment of charge transport by considering the quasiparticle description of system excitations. The notion of a quasiparticle originates from a process of adiabatic continuity, through which we establish a one-to-one correspondence between the excitations of a system of interacting particles to the excitations of a system of nearly non-interacting particles. Adiabatic continuity is a well-known concept in many-body and statistical physics, and therefore the subsequent discussion will be general and heuristic. In order to establish such a one-to-one correspondence, or to connect these systems by adiabatic continuity, we must be able to start from the excited state of the non-interacting system and turn on the interaction slowly enough so that the state occupation numbers do not change. If we consider the energy of the system and the time for this adiabatic change to be conjugate variables, then we must require that the energy ϵ of the excited state is much larger than the characteristic rate ξ at which the change occurs, i.e., $\epsilon \gg \hbar\xi$ [32].[3] This simply requires that we consider excitations at sufficiently short time scales. Furthermore, adiabatic continuity remains valid so long as the interactions do not induce transitions between states in the relevant time window, or equivalently that the lifetime of the state τ_l is long compared to the characteristic transition time, i.e., $\tau_l \gg \xi^{-1}$ [32]. Thus, we can relate the distribution functions between the interacting and noninteracting cases, supposing that we are concerned with a time scale long enough to establish a well-defined energy state but short enough so that the state

[3] In general, a factor of $\exp(-\xi t)$ controls the turning-on of the interaction.

does not decay. This requirement does not hold for strongly correlated systems. For weakly correlated systems, the quasiparticles are then the approximate eigenstates describing the excitation in the interacting system. For an electron gas, the quasiparticle distribution f satisfies the Boltzmann equation [33]. One may wonder how an equation originally used to describe weakly interacting, dilute gases can be valid for an electron gas that is hardly dilute. The resolution of this observation resides in the screening of the Coulomb interaction between electrons.

Thus within our quasiparticle framework, we define a semiclassical distribution function $f(\mathbf{r}, \mathbf{k}, t)$ to be the probability of finding a particle at $\mathbf{r}$ and at time t with momentum $\hbar\mathbf{k}$. This distribution is semiclassical because we are free to impose an uncertainty relation on $\mathbf{r}$ and $\mathbf{k}$, while we treat the force on the system classically (i.e., by Newtonian mechanics) [38]. In the absence of collisions, or alternatively for times shorter than the lifetime of the particle, we find that the total number of particles in each state $\mathbf{k}$ is conserved [32]. Thus, we can immediately write a general conservation equation for f

$$\frac{\mathrm{d}f}{\mathrm{d}t} + \nabla_{\mathbf{r}} \cdot \mathbf{j}_{\mathbf{k}} = 0 \ , \tag{2.8}$$

where $\mathbf{j}_{\mathbf{k}} = \boldsymbol{v}_{\mathbf{k}} f = (\hbar\mathbf{k}/m) f$ is the particle current in k-space under the parabolic band approximation.[4] Then, given that $\mathbf{j}_{\mathbf{k}}$ is spatially invariant (i.e., independent of $\mathbf{r}$), and taking the total derivative of f with respect to t, we have

$$\frac{\partial f}{\partial t} + \boldsymbol{v}_{\mathbf{k}} \cdot \nabla_{\mathbf{r}} f + \dot{\mathbf{k}} \cdot \nabla_{\mathbf{k}} f = 0 \ . \tag{2.9}$$

Here $\boldsymbol{v}_{\mathbf{k}} = \dot{\mathbf{r}} = (1/\hbar)\nabla_{\mathbf{k}}\epsilon(\mathbf{k})$ is the group velocity for a particle with energy $\epsilon(\mathbf{k})$, and $\dot{\mathbf{k}}$ is a classical force on the particle, which for an external electric field is simply $\dot{\mathbf{k}} = (q/\hbar)\mathbf{E}$. To account for the effect of scattering, it suffices to add a nonequilibrium term

[4] This approximation is valid for Si in the first Brillouin zone [30].

$$Q(f) \equiv \left(\frac{\partial f}{\partial t}\right)_{\text{collisions}} \tag{2.10}$$

to the right-hand-side of (2.9) [33]. We thus have the full form of the BTE in the presence of scattering

$$\frac{\partial f}{\partial t} + \boldsymbol{v}_{\mathbf{k}} \cdot \nabla_{\mathbf{r}} f + \frac{q}{\hbar}\mathbf{E} \cdot \nabla_{\mathbf{k}} f = Q(f) \ , \tag{2.11}$$

where $Q(f)$ is termed the collision integral. The left-hand-side of (2.11) is often called the streaming part of the BTE, and describes the unperturbed trajectory of a particle through the six-dimensional phase space [38]. The collision integral on the right-hand side of (2.11) captures deviations from this trajectory due to scattering. The BTE does not obey time-reversal symmetry. Thus, the processes that drive the evolution of the system are irreversible and consistent with the laws of thermal equilibrium. Indeed, consistency with the BTE requires that scattering events be numerous enough for the particles to sufficiently explore phase space, and hence the trajectories of individual particles are too complicated to be described in microscopic detail. The BTE allows us to account for the complex dynamics of individual particles by solving for their average phase space distribution, which is significantly simpler in principle.

In order to solve the BTE for a given system, we must find an explicit form for the collision integral that accounts for the scattering mechanisms present in that system. Since, in the context of this work, we are generally concerned with how charge carriers interact with the lattice, we will consider $Q(f)$ for electron-phonon interactions in the Einstein model. (In the Einstein model, we assume that each lattice ion vibrates as a harmonic oscillator independent of every other ion. This is a rough approximation to the optical phonon modes [32].) Thus, we consider a general inelastic electron-phonon scattering event in which a phonon of wave vector $\pm\mathbf{q}$ interacts with an electron of wave vector $\pm\mathbf{k}$. For the transition of a coupled electron-phonon system from state

$|\mathbf{k}\rangle \otimes |\mathbf{q}\rangle$ to state $|\mathbf{k}'\rangle \otimes |\mathbf{q}'\rangle$, we denote the transition probability per unit time as $W_{\mathbf{k}',\mathbf{k};\mathbf{q}',\mathbf{q}}$. Likewise, we denote the occupation probability for phonon state $\mathbf{q}$ as $g_{\mathbf{q}}$. Note that we are considering the general case of electron-phonon scattering, where the phonon momentum may not be fully absorbed or emitted by the electron. Then, for a discrete number of scattering events, the collision integral becomes

$$Q(f) = \sum_{\mathbf{k}',\mathbf{q}',\mathbf{q}} \Big[W_{\mathbf{k},\mathbf{k}';\mathbf{q},\mathbf{q}'} g_{\mathbf{q}'} f_{\mathbf{k}'}(1 - f_{\mathbf{k}}) - W_{\mathbf{k}',\mathbf{k};\mathbf{q}',\mathbf{q}} g_{\mathbf{q}} f_{\mathbf{k}}(1 - f_{\mathbf{k}'}) \Big] \ , \tag{2.12}$$

where we have specified the $\mathbf{k}$-dependence of f as $f_{\mathbf{k}}$ [37]. In the continuum limit of $\mathbf{k}$, the collision integral becomes

$$Q(f) = \frac{\mathcal{V}}{(2\pi)^3} \int d\mathbf{k}' \Big[\Omega_{\mathbf{k},\mathbf{k}'} f_{\mathbf{k}'}(1 - f_{\mathbf{k}}) - \Omega_{\mathbf{k}',\mathbf{k}} f_{\mathbf{k}}(1 - f_{\mathbf{k}'}) \Big] \ , \tag{2.13}$$

where

$$\begin{aligned} \Omega_{\mathbf{k}',\mathbf{k}} &= \sum_{\mathbf{q}',\mathbf{q}} \Big[W_{\mathbf{k}',\mathbf{k};\mathbf{q}',\mathbf{q}} g_{\mathbf{q}} + W_{\mathbf{k}',\mathbf{k};\mathbf{q},\mathbf{q}'} g_{\mathbf{q}'} \Big] \\ &= \frac{2\pi}{\hbar} |\langle \mathbf{k}'\mathbf{q}' | H_{el} | \mathbf{k}\mathbf{q} \rangle|^2 \delta(\epsilon(\mathbf{k}) - \epsilon(\mathbf{k}') \pm \hbar\omega_{\mathbf{q}}) \ , \end{aligned} \tag{2.14}$$

and $\mathcal{V}$ is the crystal volume. The Fermi-Dirac distribution satisfies the BTE when the collision integral is zero and, thus, describes systems that are in equilibrium. For cases where the the Fermi level lies below the conduction band in the band gap, which is typical for semiconductors, we can apply the principle of detailed balance [33]

$$\Omega_{\mathbf{k},\mathbf{k}'} f^0_{\mathbf{k}'} = \Omega_{\mathbf{k}',\mathbf{k}} f^0_{\mathbf{k}} \ , \tag{2.15}$$

where $f^0_{\mathbf{k}}$ is the Fermi-Dirac distribution. The collision integral is then

$$Q(f) = -\frac{\mathcal{V}}{(2\pi)^3} \int d\mathbf{k}' \Omega_{\mathbf{k}',\mathbf{k}'} \left[f_{\mathbf{k}} - f_{\mathbf{k}'} \frac{f^0_{\mathbf{k}}}{f^0_{\mathbf{k}'}} \right] \ , \tag{2.16}$$

We can derive a similar form for $Q(f)$ for charge scattering with impurities in the lattice, as well as for electron-electron scattering. For semiconductors at room temperature, electron-phonon scattering and impurity scattering dominate over electron-

electron scattering [38]. Indeed, the Coulomb interaction between pairs of electrons is effectively screened by other electrons, and the phase space available for scattering is limited by the Pauli exclusion principle. For devices with reduced dimensions and high carrier densities, the electron-electron contribution to scattering becomes more important, although it remains relatively weak compared to the other scattering mechanisms [37]. It is important to note that the inclusion of scattering induces a finite lifetime of the quasiparticle states. This is precisely what one should expect from inelastic process that provide a dissipative pathway for energy and momentum.

We can define a BTE for phonons as well, where we have assumed for simplicity that the only driving force is a spatial gradient, which maintains consistency with the Einstein model [32, 33]

$$\left(\frac{\partial}{\partial t} + \boldsymbol{v_q} \cdot \nabla_{\mathbf{r}}\right) g(\mathbf{r}, \mathbf{q}, t) = C(g) \ . \tag{2.17}$$

In this case, the collision integral $C(g)$ for electron-phonon scattering in the continuum limit is the same as in (2.13)

$$C(g) = \frac{\mathcal{V}}{(2\pi)^3} \int d\mathbf{k}' \Big[\Omega_{\mathbf{k},\mathbf{k}'} f_{\mathbf{k}'}(1 - f_{\mathbf{k}}) - \Omega_{\mathbf{k}',\mathbf{k}} f_{\mathbf{k}}(1 - f_{\mathbf{k}'})\Big] \ . \tag{2.18}$$

Since phonons obey Bose-Einstein statistics, the equilibrium solution to (2.17) is the Bose-Einstein distribution.

If we consider the full absorption or emission of phonons by electrons while accounting for phonon-phonon interactions with an additional collision term, then the we can write down the full form of the coupled BTE's for our electron-phonon system

$$\left(\frac{\partial}{\partial t} + \boldsymbol{v}_e(\mathbf{k}) \cdot \nabla_{\mathbf{r}} + \frac{q}{\hbar}\mathbf{E}(\mathbf{r}) \cdot \nabla_{\mathbf{k}}\right) f(\mathbf{r}, \mathbf{k}, t) = \sum_{\mathbf{q}} \left(W_{e,q}^{k+q\to k} + W_{a,-q}^{k+q\to k} - W_{e,-q}^{k\to k+q} - W_{a,q}^{k\to k+q}\right) \tag{2.19}$$

$$\left(\frac{\partial}{\partial t} + \boldsymbol{v}_p(\mathbf{q}) \cdot \nabla_{\mathbf{r}}\right) g(\mathbf{r}, \mathbf{q}, t) = \sum_{q} \left(W_{e,q}^{k+q\to k} - W_{a,q}^{k\to k+q}\right) + \left(\frac{\partial g}{\partial t}\right)_{p-p} \tag{2.20}$$

where $W_{i,\pm q}^{k+q\to k}$ denotes the emission or absorption of a phonon of wave vector $\pm\mathbf{q}$ respectively, $W_{i,\pm q}^{k\to k+q}$ denotes the emission or absorption of a phonon of wave vector $\mp\mathbf{q}$ respectively, and the final term in (2.20) is the additional phonon-phonon collision term [9]. We will further discuss phonon-phonon interactions in the next section.

We can perform a rescaling of the BTE (2.11) by introducing characteristic time, length, and velocity scales given by the average time between scattering events τ_c, the length of the device x_0, and the thermal velocity of the charge carriers $v_{\text{th}} = \sqrt{3k_B T_{\text{L}}/m^*}$, where T_{L} is the lattice temperature and m^* is the effective mass [24]. The mean free path of the charge carriers is then $\lambda_c = v_{\text{th}}\tau_c$. We introduce the dimensionless scaling parameter λ, referred to as the Knudsen number: $\lambda = \lambda_c/x_0$. The rescaled form of (2.11) is then

$$\lambda\frac{\partial f}{\partial t} + \lambda(\boldsymbol{v}_{\mathbf{k}} \cdot \nabla_{\mathbf{r}} f + \nabla_{\mathbf{r}}\phi \cdot \nabla_{\mathbf{k}} f) = Q(f) \tag{2.21}$$

where we have employed $\dot{\mathbf{k}} = -q\nabla_{\mathbf{r}}\phi(\mathbf{r})$, with the electric potential $\phi(\mathbf{r})$. This rescaling of the BTE is often called the hydrodynamic scaling. Collisions dominate a system of carriers at room temperature in a semiconductor, which implies that $\lambda \ll 1$ [24].

For lower temperatures and smaller device dimensions, the validity of the BTE becomes much more questionable due to the increasing importance of quantum mechanical effects. To be able to calculate scattering rates within the Born approximation, one must ensure that the intermediate time interval over which the distribution function evolves is much smaller than the collision time τ [37]. For a non-degenerate semiconductor, this condition can be written as

$$\frac{\hbar}{\tau} \ll k_B T \ . \tag{2.22}$$

Furthermore, when the mean free path approaches the order of the lattice spacing, $\lambda \to a$, the validity of the BTE breaks down, and one must resort to a quantum mechanical treatment of electron transport.

2.1.3 Device Equations: Moments of the Boltzmann Transport Equation

In practice, it is very difficult to obtain a solution to the BTEs (2.19) and (2.20) for a realistic semiconductor device [30]. To do this for a three-dimensional structure, one must solve for the distribution functions f and g over their complete seven-dimensional domains for a widely varying time scale[5] [9, 10]. Circumvention of this problem requires the implementation of another series of approximations. In order to simplify our system, we derive macroscopic device equations by taking weighted integrals over k-space. This procedure effectively reduces the dimensionality of the problem and is referred to as taking moments of the BTE.

For simplicity, we will consider the form of the BTE (2.11) for some general collision integral

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla_r f + \frac{q}{\hbar} \boldsymbol{E} \cdot \nabla_k f = Q(f) \ . \tag{2.23}$$

A moment is defined simply as the integration of the product of the distribution f with some weighted quantity Φ_i over the momentum space of the system

$$\langle \Phi_j \rangle \equiv \int d^3\boldsymbol{k} \Phi_j f \ . \tag{2.24}$$

A general moment of the BTE can then be written as

$$\frac{\partial}{\partial t} \langle \Phi_j \rangle + \nabla_r \cdot \langle \boldsymbol{v} \otimes \Phi_j \rangle + \frac{q}{\hbar} \boldsymbol{E} \cdot \langle \nabla_k \otimes \Phi_j \rangle = \int d^3\boldsymbol{k} \Phi_j Q(f) \ , \tag{2.25}$$

where $\otimes$ denotes a tensor product [24, 25]. Clearly, (2.25) forms an infinite series of coupled equations. This is, in fact, a compensation for the information about charge transport that is lost by integrating over momentum space. Indeed, each moment depends on quantities from higher order moments, forming an infinite hierarchy of equations that must be truncated if the problem is to be soluble. Determining how to truncate and close the system is a difficult problem for device simulation. In

[5] This variance is due to the mismatch in scattering times between optical and acoustic phonon scattering as well as impurity scattering, of which the latter possesses its own variance due to defect and doping densities.

practice, one takes the first few moments of the BTE and imposes empirically derived constitutive relations to form a closed system [38]. This will be demonstrated for the hydrodynamic model below.

To simplify the collision integral $Q(f)$, we employ the much simpler relaxation time approximation

$$Q(f) \cong -\frac{f - f_0}{\tau} \ . \tag{2.26}$$

where f_0 is the equilibrium distribution, which is simply a Fermi-Dirac distribution, and τ is a characteristic relaxation time, which is often the scattering time τ_c. The relaxation time approximation remains valid for small perturbations away from equilibrium [33]. Assuming this to be the case, the RHS of (2.25) becomes

$$\int d^3\boldsymbol{k}\Phi_j Q(f) \cong -\frac{\langle\Phi_j\rangle - \langle\Phi_j\rangle_0}{\tau_{\Phi_j}} \ , \tag{2.27}$$

where τ_{Φ_j} is defined for each local quantity Φ_j. The first four weights ($j = 0, 1, 2, 3$) typically used in practice are

$$\Phi_0 = 1 \ , \quad \Phi_1 = \boldsymbol{p} = \hbar\boldsymbol{k} \ , \quad \Phi_2 = \varepsilon = \frac{\hbar^2\boldsymbol{k}^2}{2m^*} \ , \quad \Phi_3 = \boldsymbol{v}\varepsilon \ , \tag{2.28}$$

where ϵ is the kinetic energy of the carriers [24]. Then the first four moments are

$$\langle\Phi_0\rangle = n(\boldsymbol{r}, t) \ , \quad \frac{1}{n}\langle\Phi_1\rangle = \boldsymbol{P}(\boldsymbol{r}, t) \ , \quad \frac{1}{n}\langle\Phi_2\rangle = \bar{\varepsilon}(\boldsymbol{r}, t) \ , \quad \frac{1}{n}\langle\Phi_3\rangle = \boldsymbol{S} \ , \tag{2.29}$$

where n is the carrier density, $\mathbf{P}$ is the average crystal momentum, $\bar{\epsilon}$ is the average energy, and $\mathbf{S}$ is the energy flux, or Poynting vector. Furthermore, we define $\mathbf{v} = (1/n)\langle\boldsymbol{v}\rangle$ to be the average carrier velocity. Then, referring to (2.25), the first three moment equations of the BTE are [25]

$$\frac{\partial n}{\partial t} + \nabla \cdot n\mathbf{v} = nQ_n \tag{2.30}$$

$$\frac{\partial(n\boldsymbol{P})}{\partial t} + \nabla \cdot n\langle\hbar\boldsymbol{v} \otimes \boldsymbol{k}\rangle = nQ_P \tag{2.31}$$

$$\frac{\partial(n\bar{\epsilon})}{\partial t} + \nabla \cdot n\mathbf{S} - n\mathbf{v} \cdot q\mathbf{E} = nQ_\epsilon \tag{2.32}$$

The above system constitutes a simple version of the hydrodynamic equations often employed in device simulation. Such hydrodynamic models account for energy transported by the charge carriers and the lattice. By applying the macroscopic relaxation time approximation (2.27), we can take the collision integrals to be

$$Q_n = 0 \ , \quad Q_P = -\frac{\boldsymbol{P}}{\tau_p} \ , \quad Q_\epsilon = -\frac{\overline{\epsilon} - \overline{\epsilon}_0}{\tau_\epsilon} \tag{2.33}$$

where we have introduced separate momentum and energy relaxation times and have assumed that the carrier density does not change from collision processes. Furthermore, we approximate the energy flux $\mathbf{S}$ by [24]

$$n\mathbf{S} = -\frac{5k_B T_n}{2q}\mathbf{J} - \kappa(T_n)\nabla T_n \ , \tag{2.34}$$

where

$$\mathbf{J} = \mu k_B \nabla(nT_n) + qn\mu\mathbf{E} \ . \tag{2.35}$$

Here we have introduced the carrier temperature T_n, the carrier heat capacity $\kappa(T_n)$, the carrier current $\mathbf{J}$, and the carrier mobility μ. (2.34) and (2.35) are the phenomenological constitutive relations required to close the system.[6] This model captures the fact the carrier temperature can deviate from the lattice temperature given sufficiently high electric fields and constitutes a better approximation to the BTE than drift-diffusion models. The validity of these hydrodynamic models hinges on the existence of local equilibrium. This is equivalent to the existence of one chemical potential for all electrons [33]. Thus, we implicitly assume that we can effectively model device behavior by considering the exchange of energy and momentum between subregions of the device [38].

An alternate approximation to the BTE can be made by expanding the distribution f in terms of spherical harmonics. Unlike the moment approximations, this approach does not require the assumption of local equilibrium [38]. First, we consider

[6] It has been tacitly assumed that we are working within the parabolic band approximation, which is typical for silicon devices.

a stationary form of the BTE and write the collision integral in terms of the scattering probability factors $\Omega_{\mathbf{k}',\mathbf{k}}$ defined above (2.14), where we assume that electron-phonon scattering dominates the collision processes:

$$\boldsymbol{v}_{\mathbf{k}} \cdot \nabla_{\mathbf{r}} f(\mathbf{r},\mathbf{k}) + \frac{q}{\hbar}\mathbf{E}\cdot\nabla_k f(\mathbf{r},\mathbf{k}) = \int d\mathbf{k}'\Omega(\mathbf{k}',\mathbf{k}) f(\mathbf{r},\mathbf{k}') - f(\mathbf{r},\mathbf{k}) \int d\mathbf{k}'\Omega(\mathbf{k},\mathbf{k}') \quad (2.36)$$

Here we have changed the notation, $\Omega_{\mathbf{k}',\mathbf{k}} \to \Omega(\mathbf{k}',\mathbf{k})$ and $f_{\mathbf{k}} \to f(\mathbf{r},\mathbf{k})$, for clarity. We then expand f as

$$f(\mathbf{r},\mathbf{k}) = f_0(\mathbf{r},\mathbf{k}) + \frac{k_i}{k} f_i(\mathbf{r},\mathbf{k}) + \frac{1}{2}\frac{k_i k_j}{k^2} f_{ij}(\mathbf{r},\mathbf{k}) + \dots \ , \quad (2.37)$$

where f_{ij} is a traceless tensor, $i,j = 1,2,3$, and summation over repeated indices is implied. This expansion is equivalent to a spherical harmonics expansion in momentum space, where a linear transformation relates the factors k_i/k to the first few spherical harmonics [27]. Furthermore, we assume that the scattering rates only depend on the norm of the wave vectors and expand the scattering probability as

$$\Omega(\mathbf{k}',\mathbf{k}) = \Omega_0(\mathbf{k}',\mathbf{k}) + \frac{k_i k_i'}{k^2}\Omega_1(\mathbf{k}',\mathbf{k}) + \frac{3}{2}\left(\frac{k_i k_i' k_j k_j'}{k^4} - \frac{1}{3}\right)\Omega_2(\mathbf{k}',\mathbf{k}) + \dots \ . \quad (2.38)$$

After substituting these expansions into the BTE, matching harmonic terms, truncating to the lowest two orders $i,j = 1,2$, and performing a coordinate transformation $(r,k) \to (r,H)$, where $H = E(k) - q\phi(r)$ is the total electron energy in the conduction band, we obtain the coupled set of equations

$$\frac{\partial}{\partial r_i}\left(g\lambda_1 v_k \frac{\partial f_0}{\partial r_i}\right) + 3g(H)c_{op}\Big(g^+(H)\{N_{op}^+ f_0^+(H) - N_{op} f_0(H)\} - g^-(H)\{N_{op}^+ f_0(H) - N_{op} f_0^-(H)\}\Big) = 0 \ , \quad (2.39)$$

with

$$f_i = -\lambda_1 \frac{\partial f_0}{\partial r_i} \quad (2.40)$$

Here, N_{op} and $\hbar\omega_{op}$ are respectively the optical phonon occupation number and energy, with $N_{op}^{+} = N_{op}+1$, c_{op} is a parameter related to the optical phonon coupling strength, g is the density of states with $g^{\pm}(H) = g(H \pm \hbar\omega_{op})$, and v_k is the group velocity [27]. Furthermore, $f_0^{\pm}(H) = f(H \pm \hbar\omega_{op})$, and λ_1 is the mean free path defined as above by $\lambda_1 = v_k\tau$.

In the following, we will separately employ a hydrodynamic model and a spherical harmonic expansion (SHE) to simulate charge transport in the FinFET.[7] The hydrodynamic model has been calibrated against experimental data for silicon, while the SHE model has been calibrated against 2D Monte Carlo simulations.

2.2 Phonon Scattering and Self-heating

The electron-phonon interaction is fundamental to the emergence of many significant properties of electronic materials and electronic transport, including electrical conductivity and saturation velocity. Indeed, at energy scales high enough to excite optical phonon modes, electron-phonon scattering acts as primary dissipative pathway for electron energy and momentum [8]. Likewise, phonon-phonon interactions dictate lattice thermal conductivity and lattice expansion with increasing temperature. That the origin of such macroscopic solid-state phenomena resides in quantum mechanics underscores the need to account for quantum-mechanical and semiclassical phenomena in deeply scaled transistor structures.

2.2.1 Microscopic Foundations

If we consider a lattice in which each lattice ion executes sufficiently small, harmonic oscillations around its equilibrium position (i.e., assuming the harmonic approximation), then we can quantize the normal modes of lattice vibration in their collective coordinates. These quantized modes of lattice vibration, called phonons, then constitute a collection of simple harmonic oscillators. It can be shown that

[7] As noted above, the SHE has been truncated up to second order in this work.

these phonons follow a Bose-Einstein distribution in equilibrium, thereby behaving like bosonic particles. This approach not only serves as a microscopic model for lattice heating but also provides a convenient and intuitive means of describing electron-lattice interactions via an effective field theory.

For a three-dimensional lattice, the second-quantized phonon Hamiltonian in momentum space is given by

$$H_{ph} = \sum_{\boldsymbol{k}\lambda} \hbar\omega_{\boldsymbol{k}\lambda}\left(b^{\dagger}_{\boldsymbol{k}\lambda} b_{\boldsymbol{k}\lambda} + \frac{1}{2}\right) \ , \quad [b_{\boldsymbol{k}_1\lambda_1}, b^{\dagger}_{\boldsymbol{k}_2\lambda_2}] = \delta_{\boldsymbol{k}_1,\boldsymbol{k}_2}\delta_{\lambda_1,\lambda_2} \ , \tag{2.41}$$

where $b^{\dagger}_{\boldsymbol{k}\lambda}$ and $b_{\boldsymbol{k}\lambda}$ are respectively the creation and annihilation operators for phonons of wave vector $\boldsymbol{k}$ in mode λ [32]. The displacement eigenmodes of H_{ph} are

$$\boldsymbol{u}_{\boldsymbol{k}\lambda} \equiv \ell_{\boldsymbol{k}\lambda}\frac{1}{\sqrt{s}}\left(b^{\dagger}_{-\boldsymbol{k}\lambda} + b_{\boldsymbol{k}\lambda}\right)\boldsymbol{\epsilon}_{\boldsymbol{k}\lambda} \ , \quad \ell_{\boldsymbol{k}\lambda} \equiv \sqrt{\frac{\hbar}{M\omega_{\boldsymbol{k}\lambda}}} \ , \tag{2.42}$$

where $\ell_{\boldsymbol{k}\lambda}$ is the oscillator length, $\boldsymbol{\epsilon}_{\boldsymbol{k}\lambda}$ is the polarization of the vibrational mode, and M is the mass of the lattice ion [32]. The branch index λ plays a similar role to the band index for Bloch electron states and distinguishes between optical and acoustic modes as well as labels the polarization. For n ions per unit cell, it can be shown that of the $3n$ possible modes there 3 are acoustic modes and $3(n-1)$ are optical modes [8]. For a given wave vector, $\boldsymbol{k}$, optical phonons possess an energy higher than their respective acoustic phonons. The origin of this difference is related to the relative displacements between neighboring ions. For acoustic modes, neighboring ions are displacement slightly from one another, and the displacement vector $\boldsymbol{u}_{\boldsymbol{k}\lambda}$ (2.42) maintains the same sign between neighboring sites. Conversely, for optical phonons the displacement vector alternates signs between neighboring sites, and the energy band never approaches zero. The properties of (2.41) have been used, among other things, to explain the temperature dependence of the heat capacitance of solids [33].

Similarly, we can describe the interaction between electron and phonon systems in momentum space with the second-quantized operator

$$V_{e-ph} = \frac{1}{\mathcal{V}} \sum_{\boldsymbol{k}\sigma} \sum_{\boldsymbol{q}\lambda} \sum_{\boldsymbol{G}} g_{\boldsymbol{q},\boldsymbol{G},\lambda} c^{\dagger}_{\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G},\sigma} c_{\boldsymbol{k},\sigma} \left(b_{\boldsymbol{q},\lambda} + b^{\dagger}_{-\boldsymbol{q},\lambda} \right) , \tag{2.43}$$

where $c^{\dagger}$ and c are respectively the fermion creation and annihilation operators for the electron field, $\mathcal{V}$ is the crystal volume, σ denotes the electron spin, and $\boldsymbol{G}$ gives the reciprocal lattice displacement with respect to the first Brillouin zone [32]. The interaction coupling strength g is then given by

$$g_{\boldsymbol{q},\boldsymbol{G},\lambda} = ie\sqrt{\frac{N\hbar}{2M\omega_{\boldsymbol{q}\lambda}}}(\boldsymbol{q}+\boldsymbol{G}) \cdot \boldsymbol{\epsilon}_{\boldsymbol{q}\lambda} V_{\boldsymbol{q}+\boldsymbol{G}} \tag{2.44}$$

(2.43) describes the scattering of electrons from some initial state $|\boldsymbol{k},\sigma\rangle$ to a final state $|\boldsymbol{k}+\boldsymbol{q}+\boldsymbol{G},\sigma\rangle$ via the emission of a phonon in state $|-\boldsymbol{q},\lambda\rangle$ or the absorption of a phonon in state $|\boldsymbol{q},\lambda\rangle$. Scattering process for which $\boldsymbol{G}=0$ are classified as normal processes. They tend to dominate over the Umklapp processes, for which $\boldsymbol{G} \neq 0$ [32]. For semiconductors at temperatures higher than the Debye temperature, electron scattering is dominated by optical phonons, which decay into acoustic phonons via the Klemen's channel [38]. This decay pathway originates from the phonon-phonon interaction, which involves the anharmonic terms in the phonon Hamiltonian discarded in the harmonic approximation. Equivalently, these anharmonic terms contribute to the renormalization of the phonon frequency, whereby optical phonon transition into acoustic phonons [32]. The mutual interaction between phonons then accounts for the finite lifetime of phonon modes, and the anharmonic terms are required to explain thermodynamic crystal phenomena such as the thermal expansion of solids and the lattice thermal conductivity.

Phonon dispersion curves help to elucidate the energetic properties of phonons, as well as distinguish between optical and acoustic modes. Figure 2.2 shows the theoretically predicted phonon dispersion curves in the first Brillouin zone for silicon compared against experimental values. From these dispersion curve, one can calculate the density of states and the group velocity for each mode. In particular, we

find that the higher-energy optical modes possess flatter, non-zero dispersion curves compared to the acoustic modes, and this explains their nearly-zero group velocity. The respective group velocities between acoustic and optical phonon modes, as well as the disparity in their lifetimes is central to the phonon bottleneck and the its implications for heat transport in aggressively scaled devices. We discuss this further in the next section.

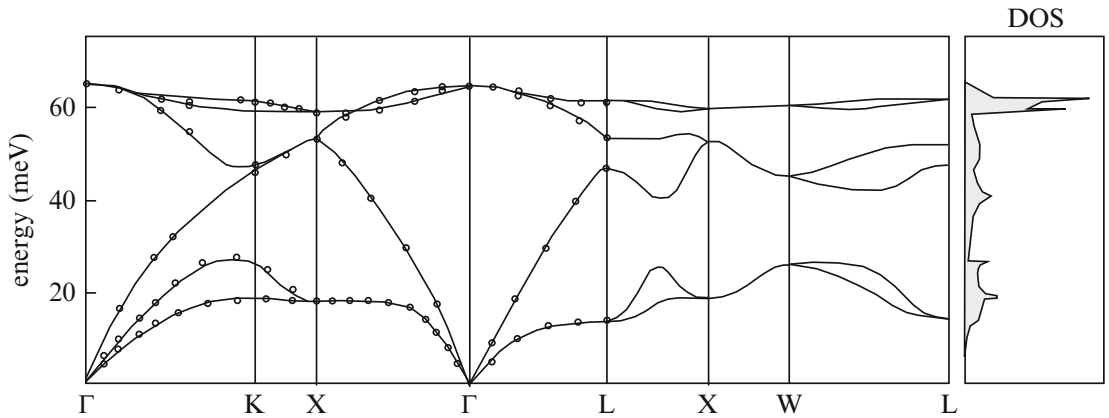


Figure 2.2: Phonon dispersion curves for silicon

2.2.2 Monte Carlo Simulation and Nonequilibrium Phonon Distributions

Monte Carlo methods provide an alternate approach to solving the BTE directly, rather than solving a closed set of macroscopic moment equations derived from the BTE. Thus one gains a more accurate simulation of sub-continuum phenomena at the cost of computational efficiency. For modeling charge transport in semiconductors, Monte Carlo methods simulate the motion of one or more charge carriers under the action of electromagnetic fields and scattering events. The scattering events and the free flight of the particle between these scattering events are selected stochastically based on a generated sequence of random numbers. The trajectory of the particle is then tracked through phase space over the course of the simulation, and then a time averaging is used to generate the distribution function f [30]. Thus, Monte

Carlo methods are better suited for determining the microscopic motion of the charge carriers. However, the complexity of semiconductor device that one can simulate is greatly limited by the high computational cost of the method [28, 29].

Much work has been done using Monte Carlo methods to model nonequilibrium transport phenomena in highly scaled two-dimensional semiconductor structures [10, 9, 14, 15, 52]. In particular, Raleva, *et al*, have addressed the phonon bottleneck problem by using a Monte Carlo algorithm to self-consistently solve the electron BTE coupled to energy balance equations for acoustic and optical phonon modes in a 2D fully-depleted SOI MOSFET with 25 nm gate length [9]. They found that electron scattering near the drain generated a significant, localized build-up of optical phonons, which caused a rise in the effective temperature of the device region. This hot spot is highly dependent on the energy of scattered electrons and the mean free path of the optical phonons. For gate and drain bias around 1.0 V, they found the temperature of the hot spot to rise 100 K higher than the equilibrium temperature of the device. Indeed these nonequilibrium phonon effects are not accounted for in drift-diffusion and hydrodynamic models.

It is important to note that obtaining a direct solution to the coupled electron and phonon BTEs, (2.19) and (2.20), even with Monte Carlo methods. Thus one must often revert to employing a moment approximation to simplify the phonon BTE (2.45):

$$\left(\frac{\partial}{\partial t} + \boldsymbol{v}_p(\mathbf{q}) \cdot \nabla_{\mathbf{r}}\right) g(\mathbf{k}, \mathbf{r}, t) = C(g) \tag{2.45}$$

The energy balance equations for acoustic and optical phonon modes can be found by taking the second-order moments of the phonon BTE and applying the relaxation time approximation [53]:

$$C_{op}\frac{\partial T_{op}}{\partial t} = \frac{3nk_B}{2}\left(\frac{T_{\mathrm{e}} - T_{op}}{\tau_{e-op}}\right) - C_{op}\left(\frac{T_{op} - T_a}{\tau_{op-a}}\right) + \frac{nm^* v_d^2}{2\tau_{e-op}} \tag{2.46}$$

$$C_a \frac{\partial T_a}{\partial t} = \nabla \cdot (\kappa_a \nabla T_a) + \frac{3nk_B}{2}\left(\frac{T_\mathrm{e} - T_\mathrm{L}}{\tau_{e-L}}\right) + C_{op}\left(\frac{T_{op} - T_a}{\tau_{op-a}}\right) \tag{2.47}$$

where T_{op} and T_a are the effective temperatures of the optical and acoustic phonon modes, C_{op} and C_a are their respective heat capacities, T_L is the lattice temperature, which is often assumed equivalent to the acoustic phonon temperature. κ_a is the thermal conductivity of the acoustic phonon system, and the τ_{i-j} are the relaxation times for phonon/lattice interaction. Furthermore, n and m^* are respectively the electron density and effective mass. We have again assumed local equilibrium for subregions of the device. This model constitutes both a semiclassical description of the Klemens' channel and an extension of the Fourier diffusion law typically employed to describe heat generation in solids.

Calculating and storing phonon energy transfer and scattering data is very computationally expensive when coupled with the BTE or moments of the BTE, such as the hydrodynamic equations [10, 11, 12]. In this work, then, we are severely restricted on the self-heating model we can employ. Thus, to obtain consistently converged solutions in a reasonably time we have implemented an empirically modified Fourier diffusion law that accounts for the heating of the lattice due to electron and hole currents and recombination phenomena.

2.3 Negative Bias Temperature Instability and the Two-stage Model

NBTI is a dominant wearout mechanism in silicon-based pMOSFET structures, originating from the build-up of positively charged defects within the SiO_2 gate dielectric and the creation of defect states at the Si/SiO_2 interface while the device is in inversion and at elevated temperatures. The aggressive scaling of device dimensions has enhanced the degradative effect of NBTI on key device parameters, namely threshold voltage V_th, drain saturation current I_{dsat} and transconductance g_m. Indeed, the increase in the importance of NBTI as a reliability concern for scaled devices can

be attributed to both the nonlinear increase in gate oxide electric fields and operating temperatures with respect to the device scale, as well as the introduction of new materials such as dual workfunction polysilicon gates (to account for short channel effects) and the use of nitrogen in the gate oxide (to control gate leakage current) [7].

NBTI is recognized as a distinct phenomena from other wearout mechanisms such as TDDB because the build-up of positive charge in the gate dielectric is not due to injection or tunneling processes. This is supported by the fact that the same degree of NBTI damage has been measured in both thick and thin gate dielectrics at similar electric fields. Indeed, the CMOS reliability community generally agrees that NBTI is caused by the cumulative contribution of two mechanisms: the generation of interface states caused by the breaking of hydrogen passivated Si-H bonds at the Si/SiO_2 interface and the generation/activation of positively charged defects in the oxide due to the trapping of holes and/or the trapping of hydrogen that has diffused into the bulk following interface bond breaking. The latter process of H diffusion is referred to as the Reaction-Diffusion mechanism. Indeed, the formation of positively charged defects in the bulk of the gate oxide has been a source of active debate and disagreement. This is largely due to significant discrepancies in the experimental literature concerning stress testing methodology and device structure and material parameters [7]. This disagreement is further complicated by the recoverable nature of NBTI degradation. Given sufficient time following an NBTI stress, one finds that degraded device parameters such threshold voltage will show a gradual return to their pre-stress values. This then requires a physical model that can consistently explain both degradation and fractional recovery phenomena. More disagreement exists as to whether recovery can be attributed to the detrapping of holes in the gate oxide or the repassivation of dangling bonds due to the backward diffusion of hydrogen.

Grasser, *et al*, have proposed a two-stage model for NBTI based on the switching behavior of neutral oxygen vacancies (E′-centers) at the Si/SiO_2 interface and the

formation of permanent, positively charged trap states (P_b-centers) [24]. This model is phenomenologically motivated by the observation of a broad scalability in NBTI data for large ranges of stress voltages and temperatures. In particular, they find that data for threshold voltage degradation versus time for various stress configurations and stress times, and for a variety of different device technologies, can be mapped onto one another by multiplication of a simple scaling factor. Accordingly, this implies that NBTI degradation is not the result of two independent dynamical processes, particularly the generation of interface states in the Reaction-Diffusion (RD) framework and the trapping of holes via elastic tunneling, as conventionally thought but, rather, by the existence of a two tightly coupled mechanisms. Indeed, they invalidate the RD framework based on the claim that it cannot accurately account for the dynamics and bias-dependence of NBTI recovery after stress. Likewise, the elastic tunneling of holes into preexisting defects fails to capture the broad scalability of the experimental data, as well as incorrectly assumes a linear dependence on the stress field and an independence to temperature.

The two-stage NBTI model proposed by Grasser, *et al*, has been shown to overcome these shortcomings [24]. This semi-empirical model is based microscopically on the trapping of holes in deep level defects (E′- and P_b-centers) via multiphonon emission (MPE) and multiphonon-field-assisted tunneling (MPFAT) processes (nonradiative capture processes). It accounts for the nonlinear-bias-field dependence and the thermal activation of NBTI degradation, as well as provides a consistent way of addressing the asymmetry between degradation and recovery times [24]. In particular, the properties of the E′- and P_b-centers justify the central role played by nonradiative capture processes in NBTI degradation.

The E′-center is a neutral oxygen vacancy (an Si-Si dimer) commonly found in amorphous SiO_2 and at Si/SiO_2 interfaces due to lattice mismatching. Experimental evidence has demonstrated that the E′-center acts as a precursor state for hole

trapping, which forms a positively charged E'_γ-center after hole trapping and lattice relaxation have occurred [35]. The defect energy level for the E′-center is typically found around 1 eV below the Si valence band, classifying it as a deep level defect state. The E'_γ-center can also act as a recombination center, capturing an electron and neutralizing the defect. That hole emission and electron capture are equivalent processes points to the switching nature of the E′-center. Thus, holes can be trapped and detrapped, and the intermediate neutral state following hole emission can relax into the initial dimer precursor state. This behavior accounts for the recoverable aspect of NBTI phenomena. The two-stage model then elevates the role of the trapping center from a purely parasitic component to the central dynamic state responsible for NBTI [24]. After a considerable number of holes have been trapped, the accumulation of positive charge enhances the creation of poorly recoverable P_b-centers via the transfer of hydrogen used to passivate dangling bonds. (P_b-centers are dangling bonds at the Si/SiO_2 interface.) The formation of poorly recoverable defects then accounts for the asymmetry between stress and recovery times. Thus, we can understand the two stages of the NBTI model: The first stage is the recoverable component and involves the trapping and detrapping of holes in neutral oxygen vacancies, while the second stage is the permanent degradation component and involves the formation of dangling bonds at the interface.

2.3.1 Microscopic Foundations: Defects and Nonradiative Capture Processes

NBTI is a thermally activated and electric-field dependent phenomena. In order to capture this dependence, a two-stage model for NBTI has been proposed based on the dynamics of oxygen vacancies (E′-centers) at the Si/SiO_2 interface [24]. Below we review the theory of nonradiative charge capture by multiphonon emission (MPE) in deep level defects (e.g. E′-centers), which is central to the two-stage NBTI model. For simplicity, we discuss the case of electron capture, while hole capture can be treated

in an analogous way.

It is generally recognized that electron capture by a defect can occur in four ways: (1) the electron transitions to the defect state by emission of a photon (photon capture), (2) inelastic electron-electron collisions cause one the electrons to lose energy and fall into the defect state (Auger capture), (3) the electron transitions through a series of closely spaced energy levels, emitting a phonon at each transition (Cascade capture), and (4) the electron emits multiple phonons as it transitions directly to the defect state (multiphonon emission capture - MPE). Capture processes (2), (3) and (4) are referred to as nonradiative transitions, as no photons are emitted during the process. For capture in deep level defects in the silicon bandgap, only multiphonon emission agrees with current experimental data [56].

A nonradiative transition via MPE requires that a free or weakly bound electronic state crosses with the bound electronic state of the defect (Figure 2.3). Such a crossing can occur for sufficiently large displacements of the lattice, as the energy level of the defect is largely dependent on the positions of its neighboring atoms in the lattice. Following the work of Huang and Rhys, we can describe the relative position of the defect level with respect to the free state with a single canonical lattice coordinate, Q. The Hamiltonian for the interacting two-level electron-lattice system is given by

$$H = H_e + H_{el} + H_l \ , \tag{2.48}$$

where H_e is the electronic Hamiltonian at $Q = 0$, H_{el} is the electron-phonon interaction Hamiltonian, describing the change in the potential well depth at Q, and H_l is the lattice Hamiltonian describing the vibrations of the lattice around $Q = 0$ [56]. Assuming linearity of the electron-phonon interaction and using s to label the vibrational modes, we can write H_{el} as

$$H_{el}(x, Q) = \sum_s u_s(x) Q_s \ , \tag{2.49}$$

where $u_s(x)$ is a linear coefficient describing the change in the potential-well depth at Q for electronic coordinate, x. Likewise, we can write H_l in the harmonic approximation as

$$H_l = \sum_s \frac{1}{2}\left(-\hbar^2 \frac{\partial^2}{\partial Q_s^2} + \omega_s^2 Q_s^2 \right) , \tag{2.50}$$

where ω_s is the phonon frequency [56, 57, 58].

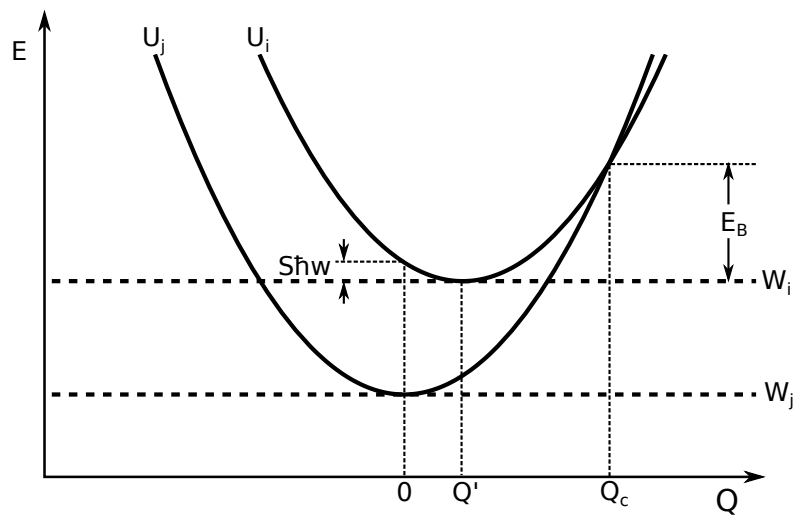


Figure 2.3: Nonradiative MPE capture process coordinate diagram

Away from level crossings, we assume that the electronic states are well-separated and, hence, follow the lattice adiabatically [56, 58]. That is, in light of the fact that the lattice ions possess a much larger mass than the free electrons, we assume that the electronic wave functions are unaffected by the motion of the lattice. Thus, we employ the adiabatic approximation to address the coupling of the electronic system to the lattice, writing the total wave function of the system as

$$\phi_{in}(x, Q) = \varphi_i(x, Q) X_{in}(Q) , \tag{2.51}$$

or alternatively in Dirac notation,

$$|in\rangle = |i\rangle|n\rangle , \tag{2.52}$$

where i denotes the electronic state and n denotes the vibrational state and vibrational quantum number. Thus, φ_i is the electronic wave function and X_{in} is the vibrational wave function respectively satisfying the following eigenfunction equations

$$[H_e(P,x) + H_{el}(x,Q)]\varphi_i(x,Q) = W_i(Q)\varphi_i(x,Q) \tag{2.53}$$

$$[H_l(P,Q) + W_i(Q)]X_{in}(Q) = E_{in}X_{in}(Q) \ . \tag{2.54}$$

Here, $W_i(Q)$ acts as an additional potential seen by the lattice, accounting for the effects of the electron-lattice interaction [57].

Note that the validity of adiabatic approximation requires that the energy of the transition is large compared to the rate of change, or equivalently when $k_bT \gg \xi$. Henry and Lang have observed that this approximation breaks down near the level crossing [56]. This can be understood conceptually by considering the behavior of the transitions as the lattice approaches the crossing point, $Q \to Q_c$. Above, we have assumed that the coupling between free state and the bound defect state is infinitesimal in order to treat the system perturbatively. However, near Q_c the electronic wave function must completely transition from a free state to a bound state within a small fraction of vibrational period. This violates the assumption of infinitesimal coupling between free and bound states. (Breakdown occurs at approximately $\epsilon \approx 0.06$ eV away from the level crossing [56].) Indeed, this breakdown in the adiabatic approximation near the level crossing is a signature of electron capture in the defect state [60].

Our point of departure from treating a general distribution of phonon modes is to assume that the electron-phonon interaction involves only a single frequency, ω_0. Then to linear order we can approximate $W_i(Q)$ as

$$W_i(Q) = W_i^0 - \sum_s (\omega_0^2 \Delta_{is}) Q_s \ , \tag{2.55}$$

where $\omega_0^2\Delta_{is}$ can be obtained by first-order perturbation theory as [57]

$$\omega_0^2 \Delta_{is} = -\int dx\ \varphi_i^*(x) u_s(x) \varphi(x) \tag{2.56}$$

Then, shifting the origin of Q_s by Δ_{is}, we introduce a change in coordinate $Q_{is} = Q_s - \Delta_{is}$, changing equation (2.54) to

$$\left[W_i + \sum_s \frac{1}{2} \left(-\hbar^2 \frac{\partial^2}{\partial Q_{is}^2} + \omega_0^2 Q_{is}^2 \right) \right] X_{in}(Q) = E_{in} X_{in}(Q)\ . \tag{2.57}$$

Here

$$W_i = W_i^0 - E_{\mathrm{LR}}^i \ \text{ and } \ E_{\mathrm{LR}} = \sum_s \frac{1}{2} (\omega_0 \Delta_{is})^2 = S\hbar\omega_0 \tag{2.58}$$

The new coordinate Q_{is} then gives the shift in the lattice equilibrium point to Δ_{is}, representing the effects of lattice relaxation [57]. Note that the relaxation of the lattice depends only on the electronic state i and reduces the energy of the system by E_{LR}^i. Here S is the Huang-Rhys factor giving the number of phonons vibrating with frequency ω_0. Thus, the lattice relaxation energy is equivalent to the phonon energy. That is, the energy lost by an electron captured in a deep level defect is carried away by a distribution of phonons. That these phonons constitute a set of independent harmonic oscillators allows us to write the eigenvalue E_{in} as

$$E_{in} = W_i + \sum_s \left(n_s + \frac{1}{2} \right) \hbar\omega_s \tag{2.59}$$

Now, implementing the Dirac notation ascribed in (2.52), we can calculate the matrix element for the nonradiative transition as

$$\langle jn' | H | in \rangle = \int dx dQ\ \varphi_j(x, Q) X_{jn'}(Q) H \varphi_i(x, Q) X_{in}(Q) \tag{2.60}$$

Based on this matrix element and using Fermi's Golden Rule, we can calculate the transition rate for electron capture with

$$\Gamma = \frac{2\pi}{\hbar} A_n^\nu \sum_{n'} |\langle jn' | H | in \rangle|^2 \delta(E_{jn'} - E_{in})\ , \tag{2.61}$$

where A_n^ν is a statistical average over the phonon distribution in some initial state i, and the summation is performed over the final phonon states n' [44, 45, 57] .

Employing the adiabatic approximation from (2.51), we find

$$\langle jn'|H|in\rangle = \int dQ X_{jn'}(Q) L_{ji}(Q) X_{in}(Q) \ , \tag{2.62}$$

where

$$\begin{aligned} L_{ji}(Q) = &- \frac{\hbar^2}{2} \sum_s \int dx \ \varphi_j(x,Q) \frac{\partial^2}{\partial Q_s^2} \varphi_i(x,Q) \\ &- \hbar^2 \sum_s \left(\int dx \ \varphi_j(x,Q) \frac{\partial}{\partial Q_s} \varphi_i(x,Q) \right) \frac{\partial}{\partial Q_s} \end{aligned} \tag{2.63}$$

is the nonadiabaticity operator [57, 58].

The zeroth order eigenvalue of H_e is given by a perturbative expansion as

$$H_e \varphi_i^0(x) = W_i^0 \varphi_i^0(x) \tag{2.64}$$

The φ_i^0 form a complete orthonormal basis $|i\rangle$ allowing us to separate diagonal and nondiagonal terms in the electron-phonon Hamiltonian H_{el} by resolution of the identity [57]:

$$\begin{aligned} H_{el} &= \sum_{i,j} |i\rangle\langle i|H_{el}|j\rangle\langle j| = \sum_i |i\rangle\langle i|H_{el}|i\rangle\langle i| + \sum_i \sum_{j\neq i} |i\rangle\langle i|H_{el}|j\rangle\langle j| \\ &= H_{elD} + H_{elND} \end{aligned} \tag{2.65}$$

Since H_{elD} does not affect the zeroth order electronic wave function φ_i^0, but rather is simultaneously an eigenfunction of $H_e + H_{elD}$ with eigenvalue $W_i^0 + \langle i|H_{el}|i\rangle$, we take H_{elND} as a perturbation to the electronic system. Thus, to first order we have

$$\varphi_i(x,Q) = \varphi_i^o(x) + \sum_k \frac{\langle k|H_{elND}|i\rangle}{W_i'(Q) - W_k'(Q)} \varphi_k^o(x) \ , \tag{2.66}$$

where

$$W_i'(Q) = W_i^0 + \langle i|H_{el}|i\rangle \ . \tag{2.67}$$

Henry and Lang have demonstrated an elegant relationship between the transition rate for MPE electron capture and the line shape for radiative capture, and they use

this to give an approximation for the capture cross section σ_c in the high temperature limit [56]. The capture cross section is related to the transition rate Γ by

$$\sigma_c = \frac{\mathcal{V}}{v_{\text{th}}}\Gamma \ , \tag{2.68}$$

where $\mathcal{V}$ is the crystal volume and v_{th} is the thermal velocity. Near the level crossing, they assume that the electronic states become independent of the lattice coordinate Q. This allows them to define the transition rate Γ for electron capture as

$$\Gamma = \frac{2\pi}{\hbar}|\langle i|\Delta V|j\rangle|^2\Big(A_n^\nu \sum_{n'} |\langle n'|n\rangle|^2 \delta(E_{jn'} - E_{in})\Big) \tag{2.69}$$

where ΔV is a small perturbation to a stationary potential, accounting for the potential $V(x, Q)$ in the electronic Hamiltonian H_e:

$$V(x, Q) = V(x, Q_0) + \Delta V(x, Q) = V_0 + \Delta V \tag{2.70}$$

They then consider, within the Condon approximation, the rate for a radiative capture process through which light is radiated with an energy $h\nu$:

$$\Gamma^{rad} = \frac{2\pi}{\hbar}\rho_{rad}|\langle i|H_{rad}|j\rangle|^2\Big(A_n^\nu \sum_{n'} |\langle n'|n\rangle|^2 \delta(E_{jn'} - E_{in} - h\nu)\Big) \tag{2.71}$$

Noting that $\rho_{rad}|\langle i|H_{rad}|j\rangle|^2 \cong (h\nu)^3$ and defining the line shape to be

$$f(h\nu) = A_n^\nu \sum_{n'} |\langle n'|n\rangle|^2 \delta(E_{jn'} - E_{in} - h\nu) \ , \tag{2.72}$$

they conclude that the transition rate for a nonradiative process is given by

$$\Gamma = \frac{2\pi}{\hbar}|\langle i|\Delta V|j\rangle|^2 f(0) \ . \tag{2.73}$$

In the limit of large S, i.e., for a system with a large number of phonons, $f(h\nu)$ can be approximated by a Gaussian:

$$f(h\nu) \cong \sqrt{\frac{1}{2\pi S(\hbar\omega)^2(2n_0+1)}}\exp\Big[-\frac{(h\nu + W_i - W_j + S\hbar\omega)^2}{2S(\hbar\omega)^2(2n_0+1)}\Big] \tag{2.74}$$

where

$$n_0 = [e^{\frac{\hbar\omega}{k_B T}} - 1]^{-1} \tag{2.75}$$

is the Bose-Einstein distribution for an equilibrium system of phonons. Then, for high temperatures $n_0 \to k_B T/\hbar\omega$, and we find that the transition rate becomes

$$\Gamma = Ae^{-\frac{E_b}{k_b T}} \ , \tag{2.76}$$

where

$$A = \sqrt{\frac{4\pi E_b}{k_B T}} \frac{|\langle i|\Delta V|j\rangle|^2}{\hbar(W_i - W_j + S\hbar\omega)} \tag{2.77}$$

and

$$E_b = \frac{(W_j - W_i - S\hbar\omega)^2}{4S\hbar\omega} \ . \tag{2.78}$$

Then, the capture cross section (2.68) is

$$\sigma_c = \frac{\mathcal{V}}{v_{\text{th}}} Ae^{-\frac{E_b}{k_b T}} \ . \tag{2.79}$$

Thus, we find a simple Arrhenius equation for the capture cross section of an MPE process for high temperatures [56, 57, 58, 61]. This exponential form has important implications for the two-stage NBTI model discussed in the next section. In particular, it accounts for the observed temperature activation of charge trapping and the associated NBTI degradation.

Nonradiative MPE transitions can also be assisted by electric fields. This can understood physically by considering that the presence of an electric field can decrease the relative separation between the conduction and defect energy levels, thereby lowering the barrier for charge capture. Ganichev, *et al*, have demonstrated that the probability for such a transition has a similar Arrhenius form as in (2.76), while acquiring a field-dependent factor $\exp(E^2/E_c^2)$, where E_c is a characteristic field value [62]. This extension of the MPE process is called the multiphonon-field-assisted tunneling (MPFAT) mechanism [24].

2.3.2 Two-stage Negative Bias Temperature Instability: The E'/P_b Switching Trap Model

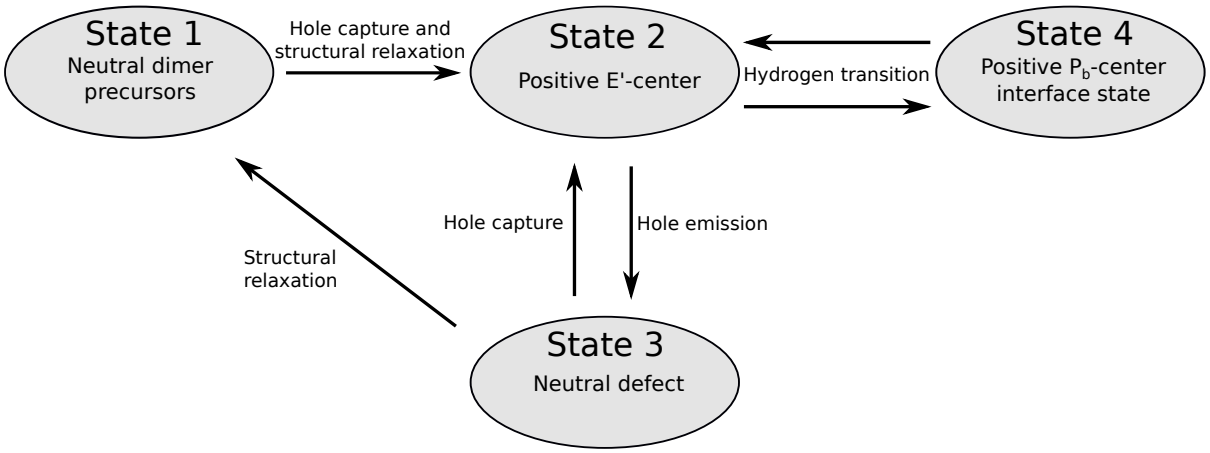


Figure 2.4: Two-stage NBTI model [39]

As discussed in the introduction to this section, we can classify four states involved in the two-stage NBTI model: (State 1) neutral oxygen vacancy, or Si-Si dimer precursor, (State 2) positively charged E'_γ-center following hole capture, (State 3) intermediate neutral defect following hole emission from State 2, and (State 4) positively charged interface, or dangling bond, state following hydrogen transfer [24]. We can model the dynamics of this four-state system by considering a coupled set of rate equations for the occupation probability f_i for a state $i = 1, 2, 3, 4$:

$$\frac{\partial f_1}{\partial t} = -f_1 k_{12} + f_3 k_{31} \tag{2.80}$$

$$\frac{\partial f_2}{\partial t} = f_1 k_{12} - f_2 k_{23} + f_3 k_{32} - f_2 k_{24} + f_4 k_{42} \tag{2.81}$$

$$\frac{\partial f_3}{\partial t} = f_2 k_{23} - f_3 k_{32} - f_3 k_{31} \tag{2.82}$$

$$\frac{\partial f_4}{\partial t} = f_2 k_{24} - f_4 k_{42} \tag{2.83}$$

Here the k_{ij} are the transition rates from state i to state j, and $\sum_i f_i = 1$. The transition rates k_{ij} for pairs of $i, j = 1, 2, 3$, with $i \neq j$, are proportional to the capture cross sections σ from the nonradiative MPE and MPFAT processes in the high temperature limit given above [56, 57]. Thus, we find a semi-empirical transition rate for holes similar to the Shockley-Read-Hall equations given by

$$k_p^c = p v_p^{th} \sigma_p e^{-x/x_{p,0}} e^{-\beta \Delta E} \theta(E_{VT}, e^{-\beta E_{VT}}, 1) e^{F^2/F_c^2} \tag{2.84}$$

$$k_p^e = p v_p^{th} \sigma_p e^{-x/x_{p,0}} e^{-\beta \Delta E} \theta(E_{VT}, e^{-\beta E_{VF}}, e^{\beta E_{TF}}) \tag{2.85}$$

where p is the hole concentration, $v_p^{th} = \sqrt{8k_bT/\pi m}$ is the hole thermal velocity, x is the distance away from the interface, $x_{p,0}$ is a characteristic hole tunneling distance, $\beta = 1/k_BT$, ΔE is the MPE energy barrier, F is the applied electric field, and F_c is the reference electric field for the MPFAT process [24]. Furthermore, we have employed the notation $E_{ij} = E_i - E_j$ and

$$\theta(E, \mu, \nu) = \begin{cases} \mu & E \geq 0 \\ \nu & E < 0 \end{cases} \tag{2.86}$$

Then E_F, E_V, and E_C are respectively the Fermi level in the channel, the valence band at the interface, and the conduction band at the interface. For Si/SiO$_2$, $\sigma_p \approx 3 \times 10^{14}$ cm^2, $x_{p,0} \approx 0.05$ nm, and $F_c \approx 2.83 \times 10^6$ V/cm. Using the notation k(trap level, MPE barrier ΔE, MPFAT reference field), we can then write the transition rates k_{ij} between the states $i, j = 1, 2, 3$ as

$$k_{12} = k_p^c(E_T, \Delta E_B, F_C) \tag{2.87}$$

$$k_{23} = k_p^e(E_T', \Delta E_C, 0) \tag{2.88}$$

$$k_{32} = k_p^c(E_T', \Delta E_C, F_C) \tag{2.89}$$

$$k_{31} = \nu_1 e^{-\beta \Delta E_A} \tag{2.90}$$

where $\nu_1 \approx 10^{13}$ Hz is the attempt frequency for thermally-activated transitions over energetic barriers [24]. E_T is the energy level of the trap state 1, and E_T' is the energy level for trap states 2 and 3, which is distinct from E_T due to lattice relaxation.

Furthermore, ΔE_A, ΔE_B, and ΔE_C are the MPE barriers associated with state 3-1 transitions, state 1-2 transitions, and state 2-3 transitions respectively. We likewise assume that transitions between states 2 and 4 are thermally activated and thus given by the Arrhenius type equations:

$$k_{24} = \nu_2 e^{-\beta(\Delta E_D - E_2 - \gamma F)} \tag{2.91}$$
$$k_{42} = \nu_2 e^{-\beta(\Delta E_D - E_4 - \gamma F)} \tag{2.92}$$

Here $\nu_2 \approx 5.11 \times 10^{13}$ Hz is the attempt frequency and ΔE_D is the MPE barrier for state 2-4 transitions, E_2 and E_4 are the trap levels for states 2 and 4, respectively, and $\gamma \approx 7.4 \times 10^{-8}$ cm/V.

In order to obtain macroscopic quantities of interest, in this case the number of holes trapped at the interface N_{it} and in the oxide N_{ox} as well as the total charge trapped at the interface Q_{it} and in the oxide Q_{ox}, we take moments of the state occupation probabilities f_i, in principle similar to the moments method employed in deriving macroscopic equations from the BTE. Thus, we have

$$N_{\text{it}}(t) = \langle f_4(t) \rangle \ , \tag{2.93}$$
$$N_{\text{ox}}(t) = 1 - \langle f_1(t) \rangle \ , \tag{2.94}$$
$$Q_{\text{it}}(t) = q\langle f_4(t)(1 - f_{\text{it}}(t) \rangle \ , \tag{2.95}$$
$$Q_{\text{ox}}(t) = q\langle (1 - x/t_{\text{ox}})(f_2(t) + f_4(t)) \rangle \ , \tag{2.96}$$

where the moment is defined as

$$\langle f \rangle = N \int \mathrm{d}\mathbf{\Omega} f g(\mathbf{\Omega}) \ . \tag{2.97}$$

Here N is the number of defects, $g(\mathbf{\Omega})$ is a weight function giving the joint probability density, and we have averaged over the set of random variables $\mathbf{\Omega}$:

$$\mathbf{\Omega} = (x, E_T, E_T', \Delta E_A, \Delta E_B, \Delta E_C, \Delta E_D) \ . \tag{2.98}$$

Each defect is then defined by a unique configuration of $\mathbf{\Omega}$ [24]. The change in Q_{it} and Q_{ox} over time,

$$\Delta Q_\mu(t) = Q_\mu(t) - Q_\mu(0) \ , \tag{2.99}$$

gives the shift in threshold voltage over time $\Delta V_{\rm th}(t)$ according to

$$\Delta V_{\rm th}(t) = -\frac{\Delta Q_{\rm ox}(t) + \Delta Q_{\rm it}(t)}{C_{\rm ox}} \ , \tag{2.100}$$

where $C_{\rm ox}$ is the capacitance of the gate oxide. Often one measures the change in the drain current I with respect to a reference time $t_{\rm ref}$ to extract the shift in threshold voltage [7]:

$$\Delta V_{\rm th}(t) = -\frac{I(t) - I(t_{\rm ref})}{g_m(t)} \ , \tag{2.101}$$

where $g_m(t) = \partial I/\partial V$ is the transconductance extracted from I-V measurements. This is the commonly employed on-the-fly method for NBTI measurements. We will employ this method to calculate $\Delta V_{\rm th}(t)$ from an NBTI simulation utilizing the two-stage model.

2.3.3 The Reaction-Diffusion Framework and Unified Reliability Models

We have chosen the two-stage NBTI model over a reaction-diffusion model because of its direct generalizability to more accurate phonon interactions and scattering processes not available in a typical RD framework. Not only does the nonradiative capture by MPE process serve as the foundation for two-stage NBTI, it also provides a consistent framework for understating $1/f$ noise and, more recently, TDDB in high-k gate dielectrics [42, 43]. Indeed, such an approach lends itself to a more unified way of thinking about reliability phenomena and may provide an avenue for predictive reliability analysis in novel nanostructures.

There are a few possible limitations to the two-stage model. According to Mahapatra, *et al*, the long-time behavior of NBTI degradation has not been predicted or verified with the two-stage model, nor with other, similar well-based models [21]. Indeed, it has been shown that long-time behavior of NBTI degradation is domi-

nated by H diffusion in the dielectric, which the R-D framework accurately captures. Furthermore, the two-stage model has failed to accurately reproduce experimental signatures for the duty cycle and frequency dependence of NBTI degradation under AC stress. This is due, in part, to the calibration of adjustable parameters in the two-stage model, which the RD framework does not rely on. Finally, the inclusion of the effects of process variability on NBTI degradation is not explicit for the two-stage model, whereas the RD framework can phenomenologically accommodate such effects. Nonetheless, it possible to incorporate the physics of hole trapping described by the two-stage model into the R-D framework to understand the contributions of deep-level defects and H diffusion during NBTI stress [21].

2.4 The Finite-Element Simulation Environment and Negative Bias Temperature Instability Stressing

For this work, we have employed Sentaurus TCAD to solve a coupled set of hydrodynamic and thermodynamic device transport equations along with the two-stage NBTI model for a two-dimensional and three-dimensional FinFET structure. In general, TCAD simulation uses a finite element routine to solve a discretized set of device equations on a refined finite element mesh. In particular, the Senaturus TCAD structure is multi-functional, and has been used to build the geometric structure of the FinFET devices with prescribed material parameters, to generate an appropriate mesh, and to solve for two separate transport equation sets accounting for NBTI. In the following, we will give an overview of this simulation environment, emphasizing the structure of the NBTI simulation.

The finite element mesh is generated by a Delaunay triangulation algorithm. This algorithm is particularly suited to handling corners and material boundaries in device structures. An example mesh is given in Figure 2.5 for the 2D FinFET structure will consider in this work. Here, the mesh has been refined such that a high concentration of nodes exists near material boundaries and at regions of high charge transport.

Information about material parameters and boundaries is contained on the discrete mesh, and continuous solutions are obtained by interpolation methods.

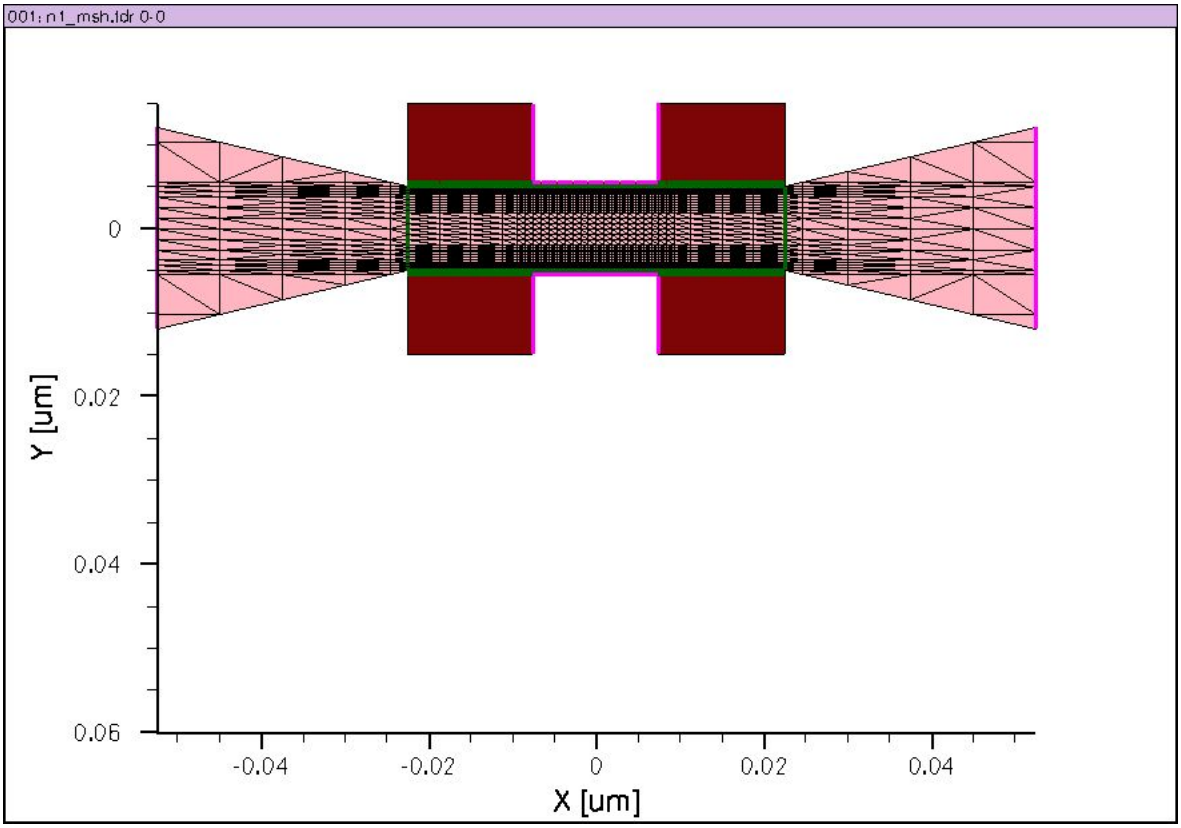


Figure 2.5: Example finite element mesh for 2D FinFET structure

SDevice is the primary simulation component of Sentaurus TCAD that has been employed in this work, and provides a variety of numerical solvers based on the discretized device structure. SDevice, then, is used to find self-consistent solutions to a variety of device phenomena, including NBTI. For this work, we use SDevice to solve, separately, the hydrodynamic and SHE models coupled to the two-stage NBTI model and a self-heating equation. This heating equation is an empirically modified Fourier diffusion equation that accounts for the increase in lattice temperature due to electron and hole currents as well as generation and recombination phenomena.

The programs TecPlot and Inspect, which are part of the Sentaurus package, have been used to render device images and data plots, respectively.

2.4.1 Simulating Negative Bias Temperature Instability

The NBTI simulation consists of three parts run successively with SDevice. These parts will be referred to as the prestress, the stress, and the relaxation phases of the simulation. In the prestress phase, a quasistationary simulation[8] is run in which the gate and drain are ramped to their initial values, $V_{\mathrm{g}} = -0.3$ V and $V_d = -0.05$ V, and the initial density of precursor trap states (state 1) is set, $N_0 = 5.0 \times 10^{12}$ cm^{-2}. A transient simulation is then run which characterizes the transconductance g_m of the device. A second transient simulation follows, which evolves the NBTI degradation of the system for 10^{-3} s, and the data from this last step sets the initial conditions for the stress phase.

During the stress phase, a quasistationary simulation ramps the gate and drain biases and the ambient device temperature to their stress condition values. For this work, we consider two stress configurations. The first is a symmetric stress, in which the source is grounded, the drain is held at a low, near-zero negative voltage (i.e., the prestress value), and the gate is set to a larger negative voltage. This is the standard stress configuration for measuring NBTI degradation. The second configuration is asymmetric, in which the drain is ramped to the same bias as the gate over the course of the stress. This stress configuration is typical in analog ICs. Following the quasitationary ramp, a transient simulation runs over the stress time and solves for the two-stage NBTI trap states coupled to the device equations. Thus, the transient simulation accounts for NBTI degradation of the device parameters. For this work, we have considered the short-time behavior of NBTI degradation. Thus the transient

[8]Sentaurus TCAD distinguishes between two types of simulation which can be run under the same command file. A quasistationary simulation solves the coupled device equations while iterating through a virtual time parameter. During such a simulation, the device is brought to a steady-state according to specified electrical and thermal boundary conditions. The quasistationary simulation is often used to set the initial conditions of the second type of simulation, a transient simulation. During a transient simulation, the state of the device is allowed to evolve with time for a given set of boundary conditions. Thus, the simulation solves the coupled device equations while iterating through real time.

stress runs for 1 s. The transient simulation data is then read into the relaxation phase of the NBTI simulation.

During the relaxation phase, the gate and drain are ramped back down to their prestress values, and a transient simulation calculates the fractional recovery of NBTI degradation as the trap states relax. This transient simulation is run for 10^3 s to capture the asymmetry in degradation and recovery times. A schematic representation of this simulation flow is given in Figure 2.6.

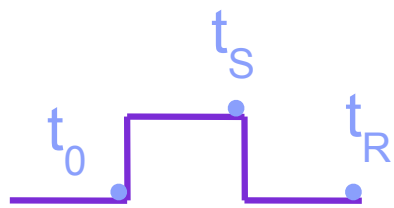


Figure 2.6: Heuristic NBTI simulation steps

We implement two transport models: the hydrodynamic equations and the SHE of the Boltzmann transport equation. These are coupled to the Poisson equation, the thermodynamic self-heating equation, and the two-stage NBTI model. The SHE serves as a better approximation to the high energy tail of the charge carrier distribution and has been calibrated for silicon against Monte Carlo simulation. Its computational cost is far higher than that of the hydrodynamic equations, and thus the SHE is only implemented for 2D devices. For 3D devices, we resort to the hydrodynamic model.

Each NBTI simulation records the steady-state and transient solutions to the device equations as well as the occupation of the trap states. After each simulation, the shift in threshold voltage V_{th} is determined by the on-the-fly method discussed above. Thus, $V_{\text{th}}(t)$ is calculated using the transconductance and the change in drain current according to equation (2.101).

Table 2.1: Comparison between hydrodynamic and SHE models

	Strengths	**Weaknesses**
SHE	• Better captures high energy tail of hole distribution (calibrated against Monte Carlo simulation) • Directly compatible with two-stage NBTI model • Explicit spacial dependence	• More computationally intensive • Not yet compatible with 3D simulations
Hydrodynamic	• Less computationally intensive • Directly compatible with two-stage NBTI model	• Assumes more approximate dynamics of charge transport • Calibration cannot be as finely tuned

For the two-stage NBTI model, we need only specify the range for the independent random variables characterizing the defect. The precursor state energy level E_T and the E′-center trap state energy level E'_T follow uniform distributions, with $-1.14 \leq E_T \leq -0.31$ eV and $0.01 \leq E'_T \leq 0.3$ eV. The P_b-center trap state energy level follows a Fermi-Dirac distribution, with an average of 0.5 eV and a standard deviation of 0.44 eV. These are measured with respect to the valence band edge. For simplicity, we take $\Delta E_C \ll \Delta E_B$ and approximate $\Delta E_A \approx \Delta E_B$, with $0.01 \leq \Delta E_A \leq 1.15$ eV [39].

CHAPTER 3

SIMULATIONS AND RESULTS

In the following, we examine three case structures: an idealized 2-dimensional pFinFET with gate length $L_g = 15$ nm, a 3-dimensional pFinFET with $L_g = 15$ nm, and a 2-dimensional pMOSFET with $L_g = 40$ nm. In particular, we focus on the 2D pFinFET for a primary qualitative characterization of NBTI and self-heating behavior, using the 2D pMOSFET as a check for consistency. We then test for the same NBTI and self-heating phenomena on the 3D pFinFET, again noting the qualitative behavior and differences between each simulation. In most cases, we have employed Dirichlet thermal boundary conditions, where the thermal boundaries correspond to the electrical contacts of the device. Dirichlet boundary conditions assume that heat is rapidly transported out of the device through the contacts, or equivalently that the contacts have near-zero thermal resistance. For the 2D FinFET case, we will also briefly examine the affects of Neumann conditions, where we allow the contact temperature to change in response to self-heating.

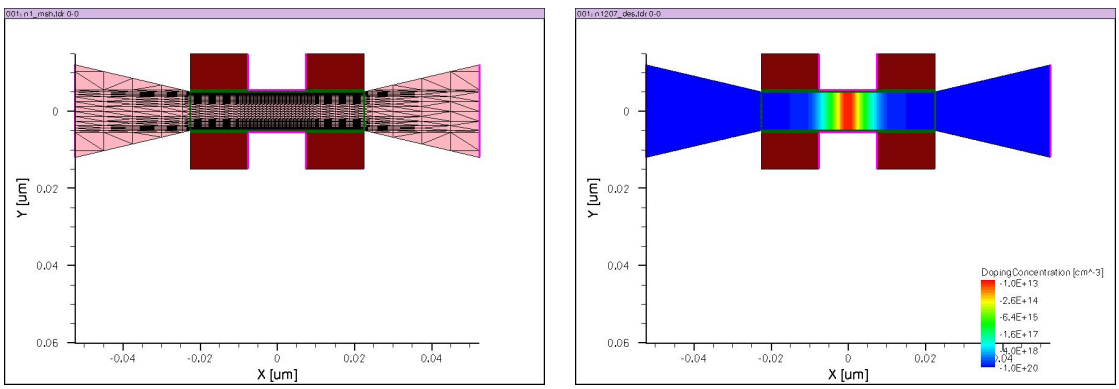
Table 3.1: Simulation cases

	Stress Configuration		**Thermal BCs**		**Transport Model**	
	Symmetric	Asymmetric	Dirichlet	Neumann	Hydrodynamic	SHE
2D MOSFET	**X**		**X**			**X**
2D FinFET	**X**	**X**	**X**	**X**	**X**	**X**
3D FinFET	**X**		**X**		**X**	

3.1 Case 1: 2D pFinFET

The 2D FinFET consists of a long, narrow p-doped, $\langle 100 \rangle$-oriented, Si channel of length $L_C = 40$ nm surrounded on both sides by aligned gates of length $L_g = 15$

nm with an SiO_2 gate dielectric of thickness $t_{\text{ox}} = 2$ nm. The gates are separated from extended source and drain regions by oxynitride spacers. Thus, the 2D FinFET constitutes a typical dual-gate structure. The basic TCAD structure with finite element mesh is shown in Figure Figure 3.1(a). We have applied a gaussian doping concentration centered in the channel using boron as the dopant. This is shown in Figure 3.1(b). Electrical contacts / thermal boundaries at source, drain, and gates are indicated by the pink edges, the gate material itself has been removed and replaced with the contact edge. This is a common feature of TCAD devices.



(a) Finite element mesh (b) Doping concentration

Figure 3.1: 2D FinFET structure

We present the results for two stress configurations for a range of gate voltages V_{g} and ambient temperatures T_{ext}. The first stress configuration is symmetric, where the source has been grounded, the drain is held at a near-zero bias ($V_d \approx -0.05$ V), and the gate is raised to a large negative bias. The second stress configuration is asymmetric, where the source is grounded and the gate and drain are raised to the same negative bias ($V_{\text{g}} = V_d$). This simulation employs an SHE model (as opposed to a hydrodynamic model) as part of the system of transport equations used to describe the operation of the device.

3.1.1 Symmetric Stress Configuration

We examine the degradation of threshold voltage over time as the device is stressed in the short-time regime (t_{stress} = 1 s), which serves as the main indicator of NBTI wearout in the device. In general, we test at combinations of ambient temperature $T_{\text{ext}} = \{300, 325, 350, 375\}$ K and gate voltage V_{g} = {-0.5, -0.75, -1.0, -1.25, -1.5, -1.75, -2.0} V. First, we fix the gate voltage and vary the ambient temperature. The results are shown in Figure 3.2, Figure 3.3, and Figure 3.4. We find that the rate of V_{th} degradation increases monotonically with respect to increasing temperature, exhibiting a logarithmic time dependence as expected in the short-time regime [24]. This monotonic behavior is expected given the Arrhenius form of temperature dependence for charge trapping associated with NBTI.

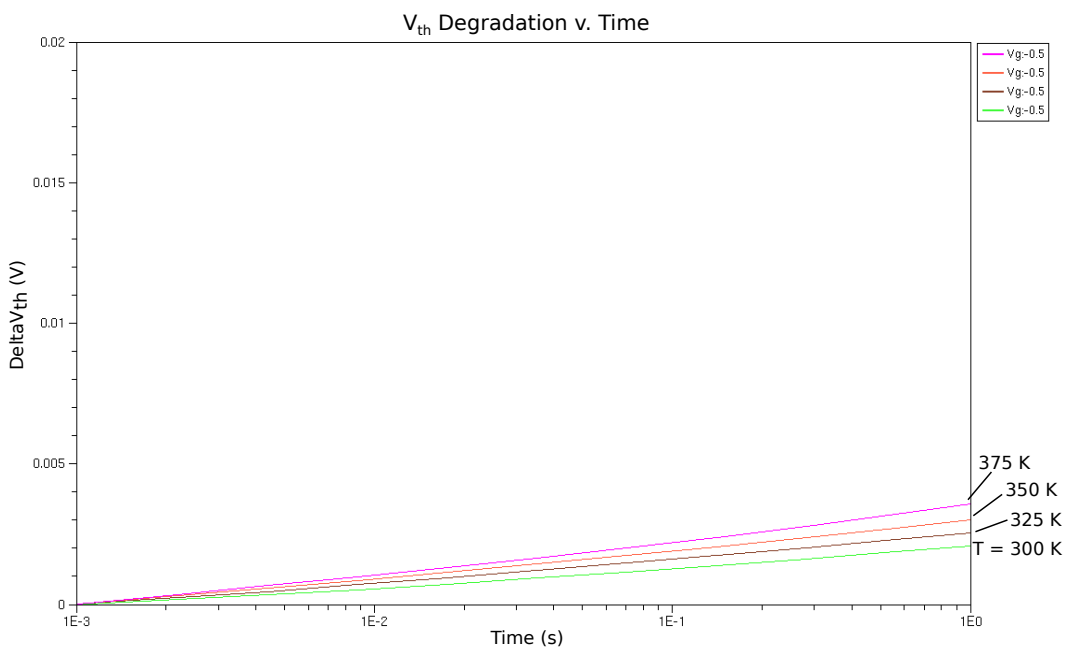


Figure 3.2: Temperature variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $V_{\text{g}} = -0.5$ V

Next, we fix the ambient temperature and examine $\Delta V_{\text{th}}(t)$ for different gate voltages. The results are shown in Figure 3.5 to Figure 3.13.

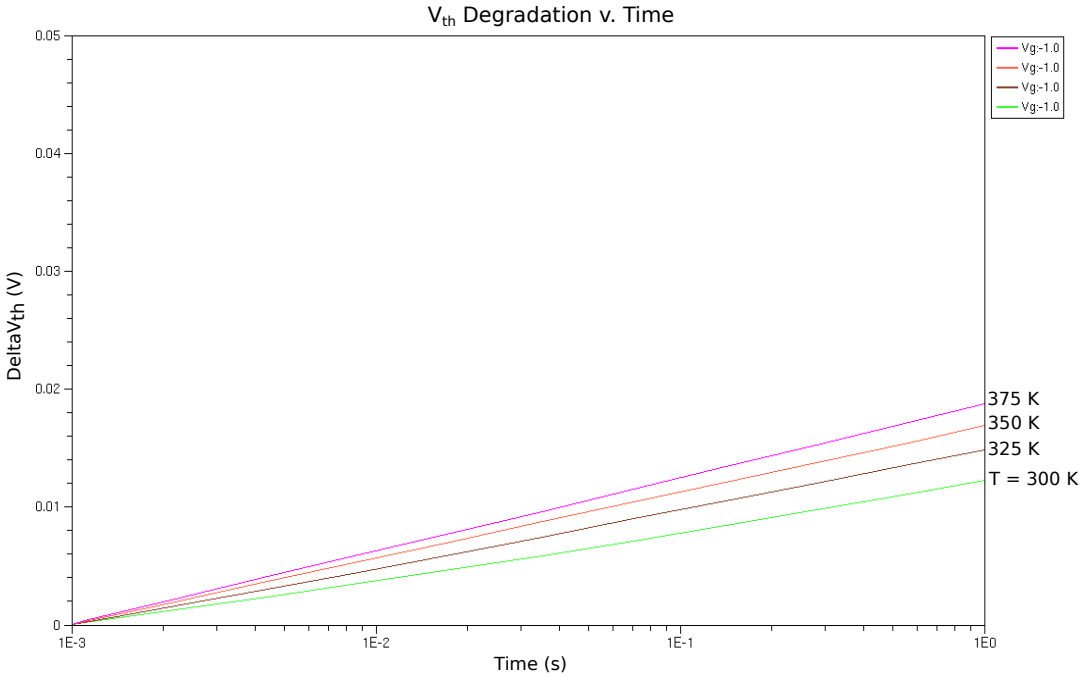


Figure 3.3: Temperature variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $V_{\mathrm{g}} = -1.0$ V

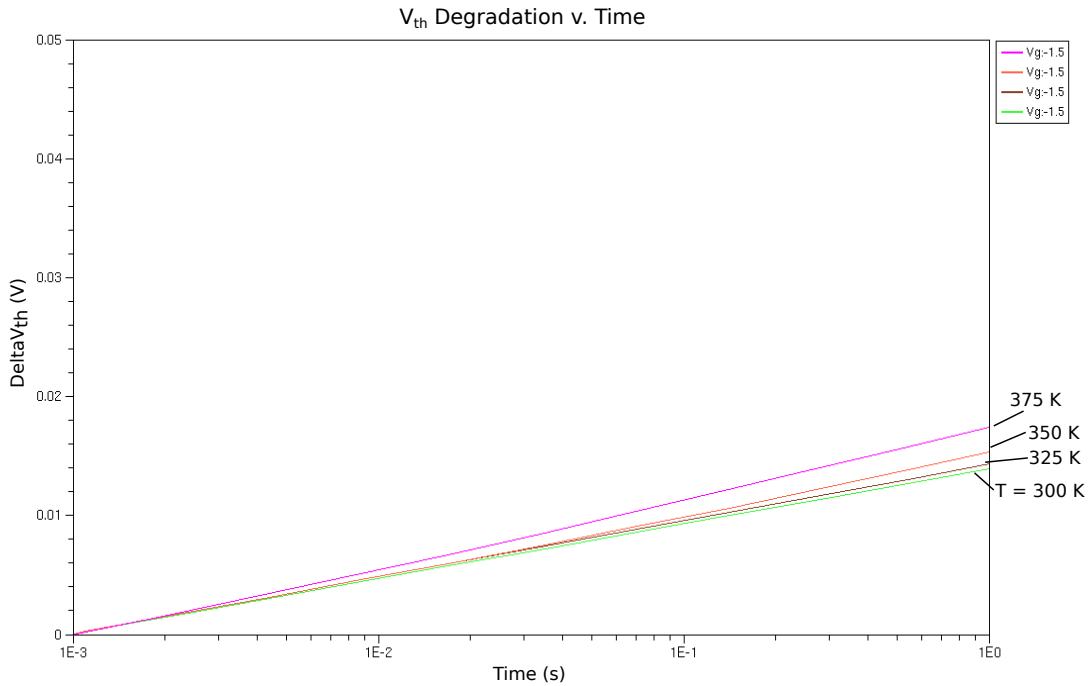


Figure 3.4: Temperature variation of $\Delta V_{\mathrm{th}}(t)$ during symmetric stress, $V_{\mathrm{g}} = -1.5$ V

We find that at lower temperatures the degradation again follows a logarithmic time dependence as expected. However, the rate of V_{th} degradation with respect to gate voltage no longer exhibits monotonically increasing behavior at lower temperatures. That is, for $T_{\text{ext}} = 325$, 350, 375, 400 K, the ΔV_{th} curves increase in slope up to around -1.0 V, at which point the monotonic behavior is broken and the rate of degradation either improves or oscillates. As the temperature is raised beyond $T_{\text{ext}} = 450$ K, the monotonic behavior is restored, and the V_{th} curves begin to deviate from a logarithmic time dependence. We summarize this broken monotonic behavior for increasing gate voltage under symmetric stress in Figure 3.14. Furthermore, we find that the monotonic behavior of the degradation rate is also restored under asymmetric stress configurations. Figure 3.15 shows this for $T_{\text{ext}} = 350$ K, where we observe that the range in the magnitude of V_{th} is much larger compared to the symmetric stress configuration.

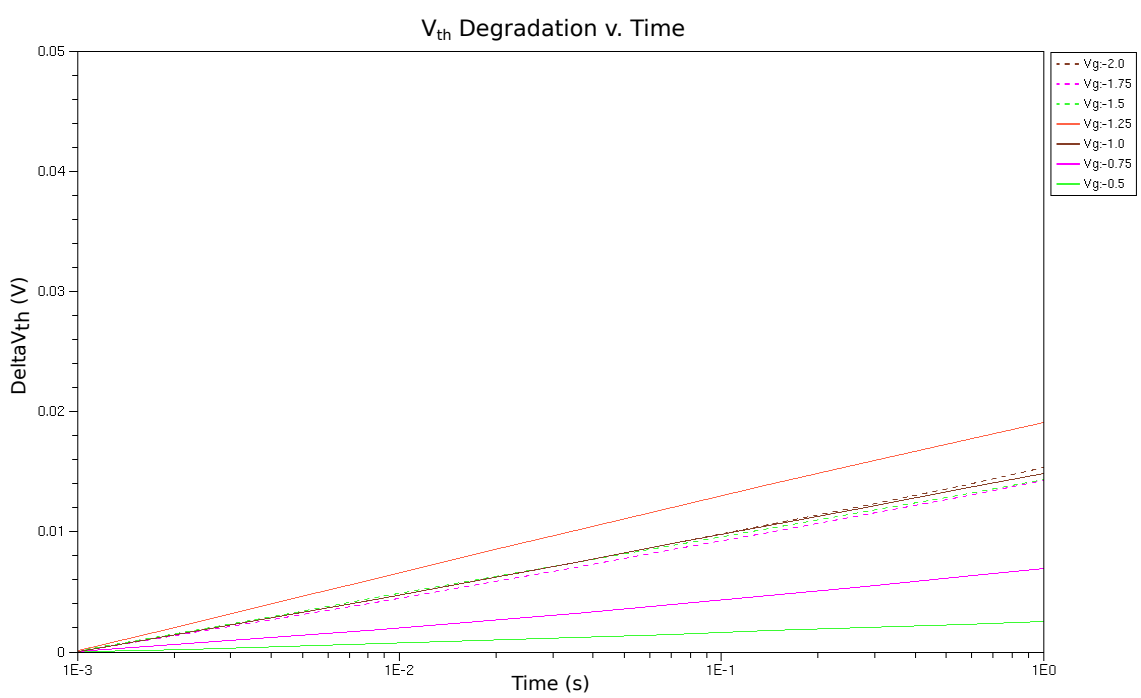


Figure 3.5: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 325$ K

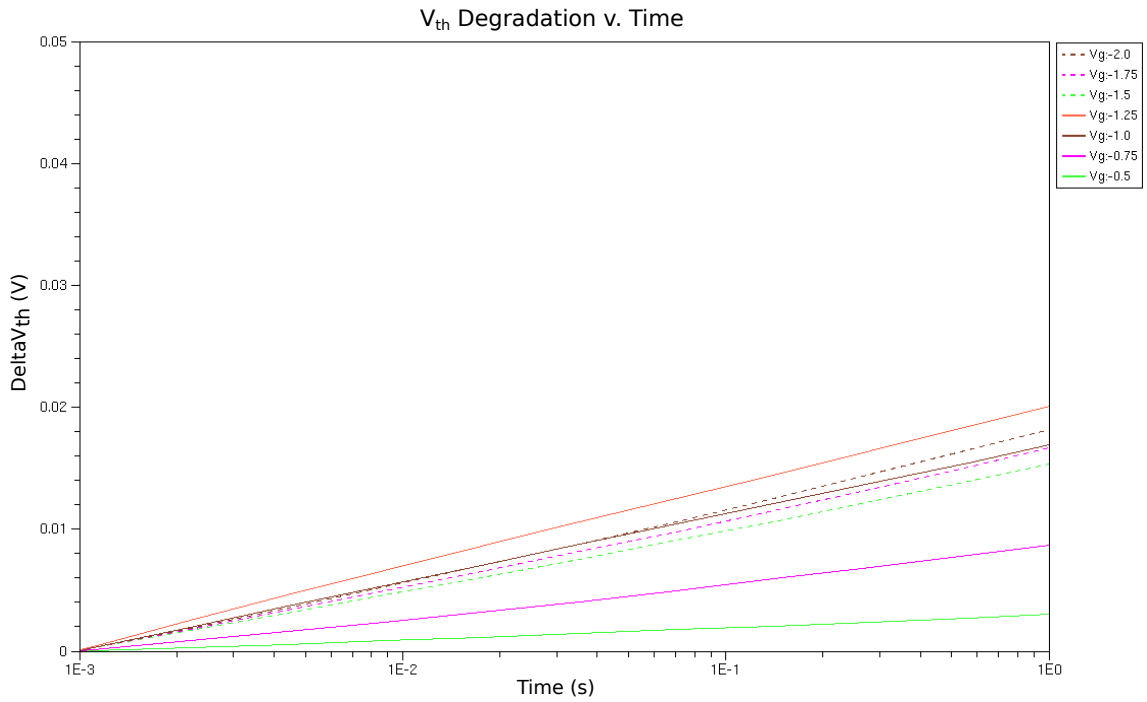


Figure 3.6: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 350$ K

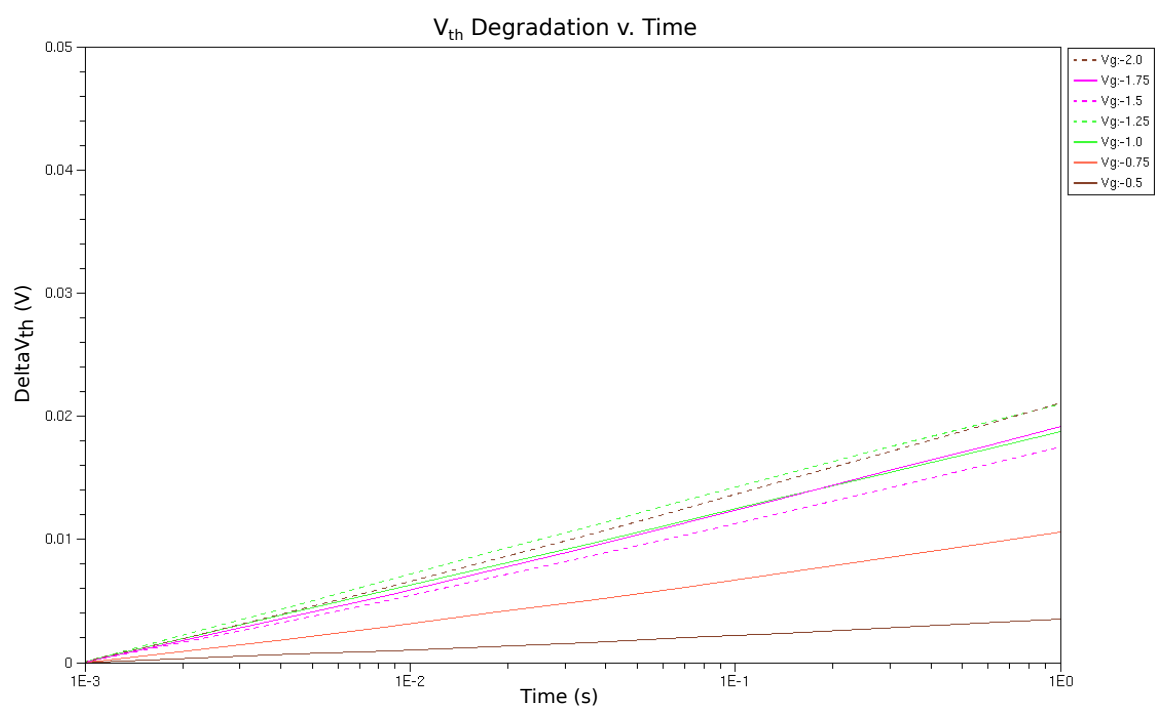


Figure 3.7: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 375$ K

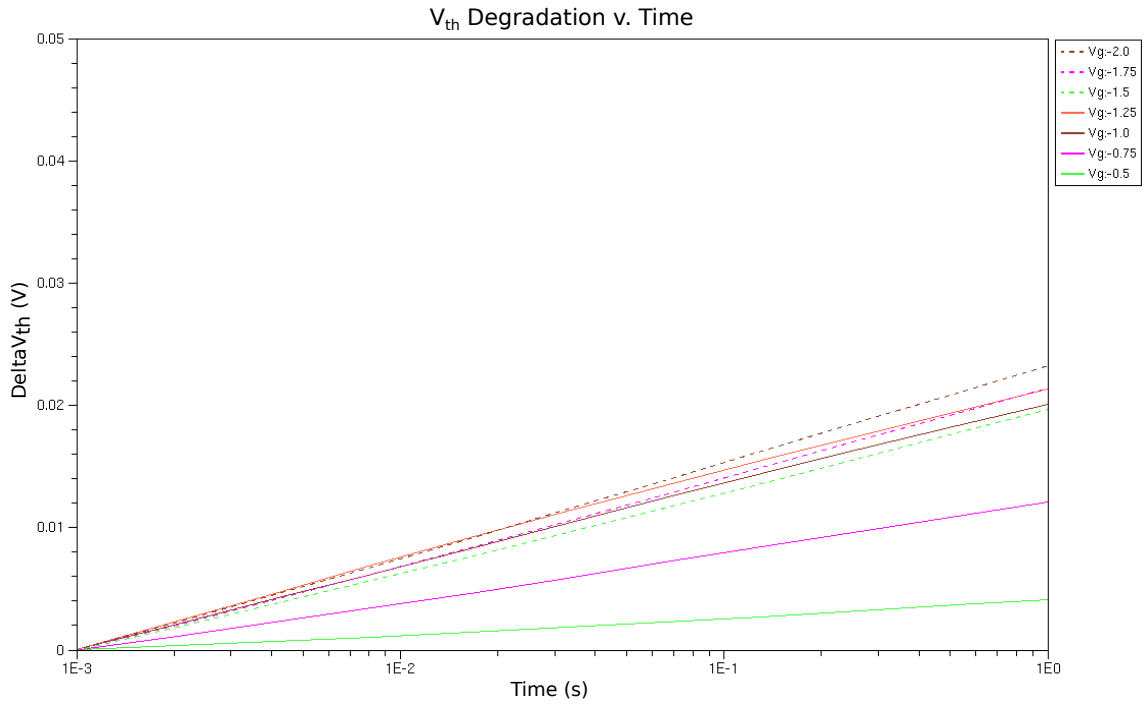


Figure 3.8: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 400$ K

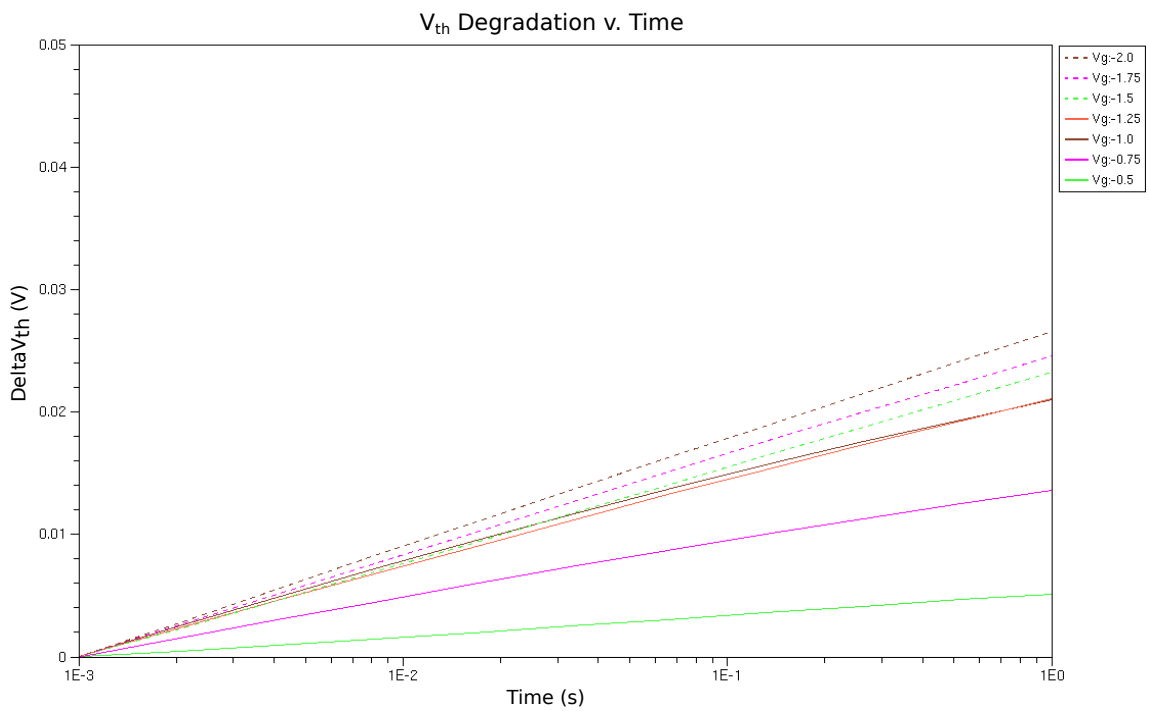


Figure 3.9: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 450$ K

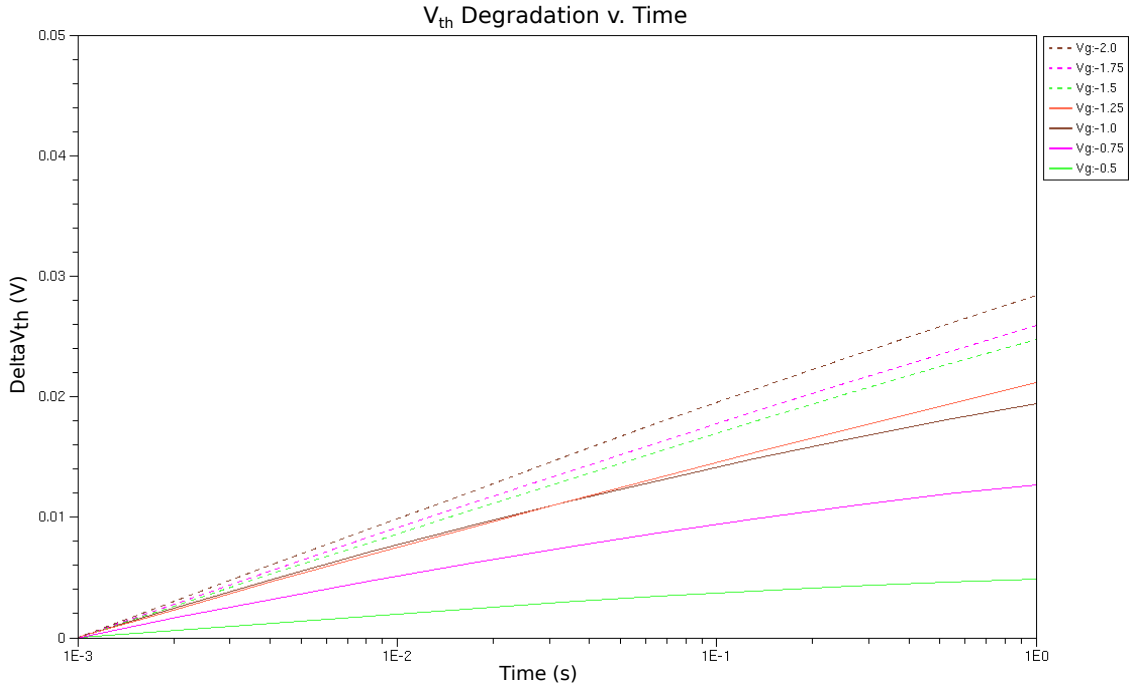


Figure 3.10: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 500$ K

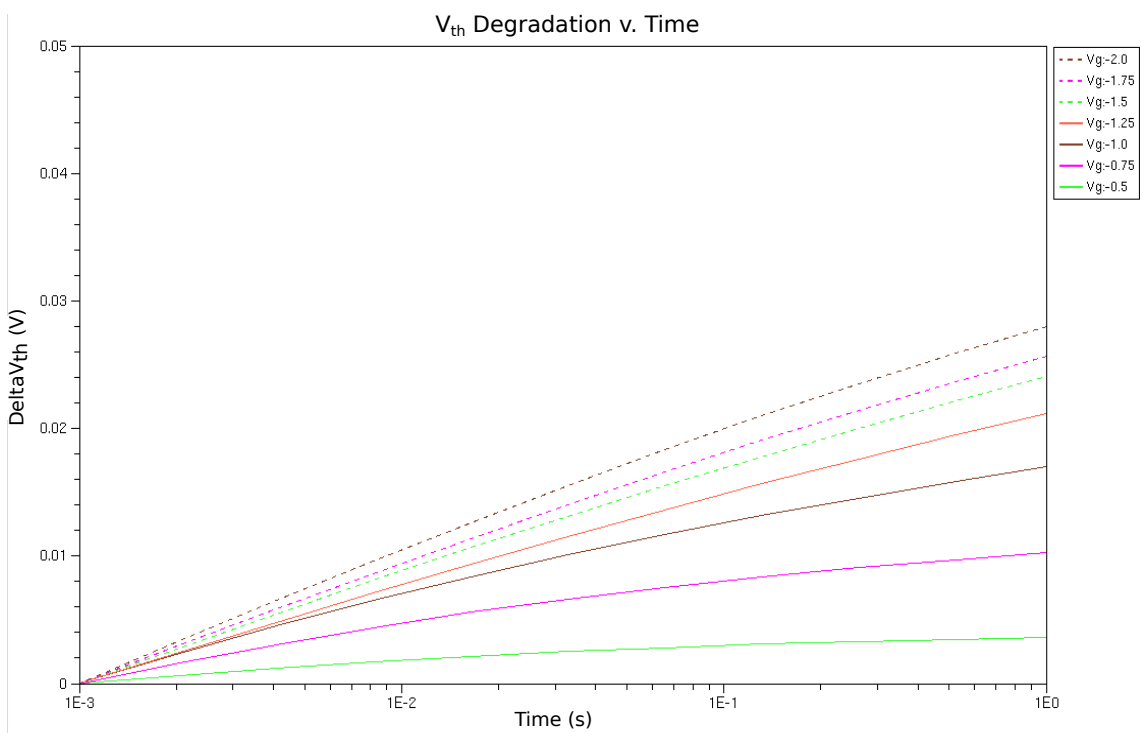


Figure 3.11: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 550$ K

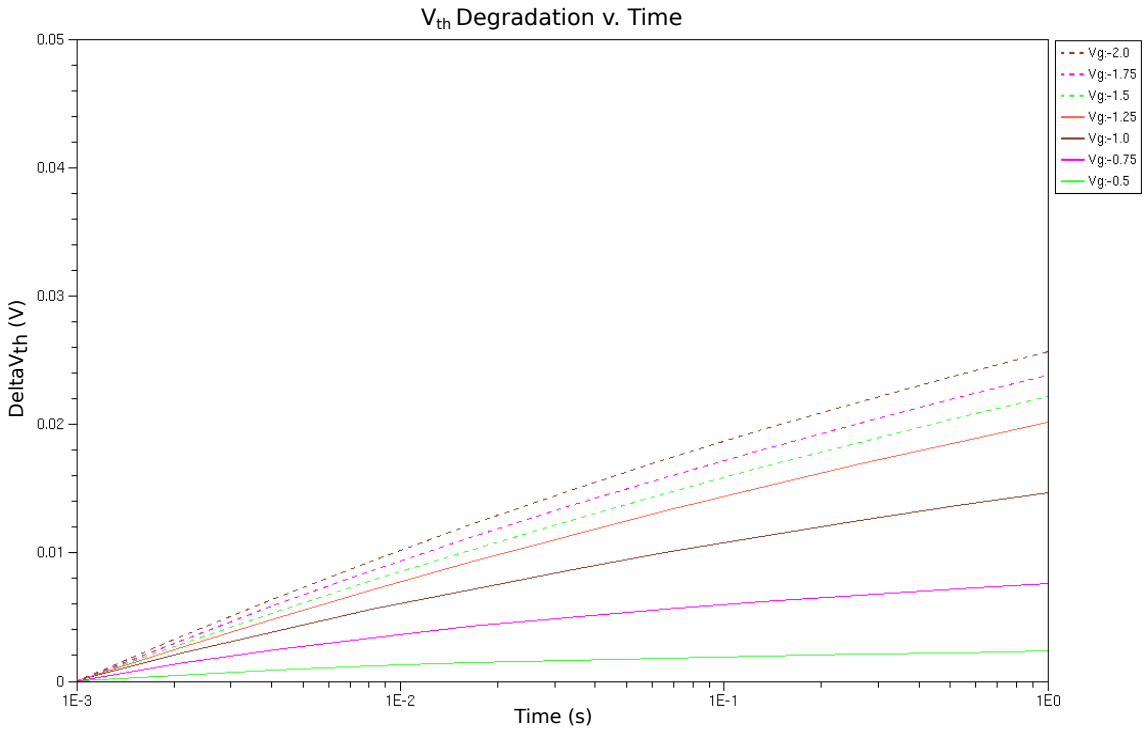


Figure 3.12: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 600$ K

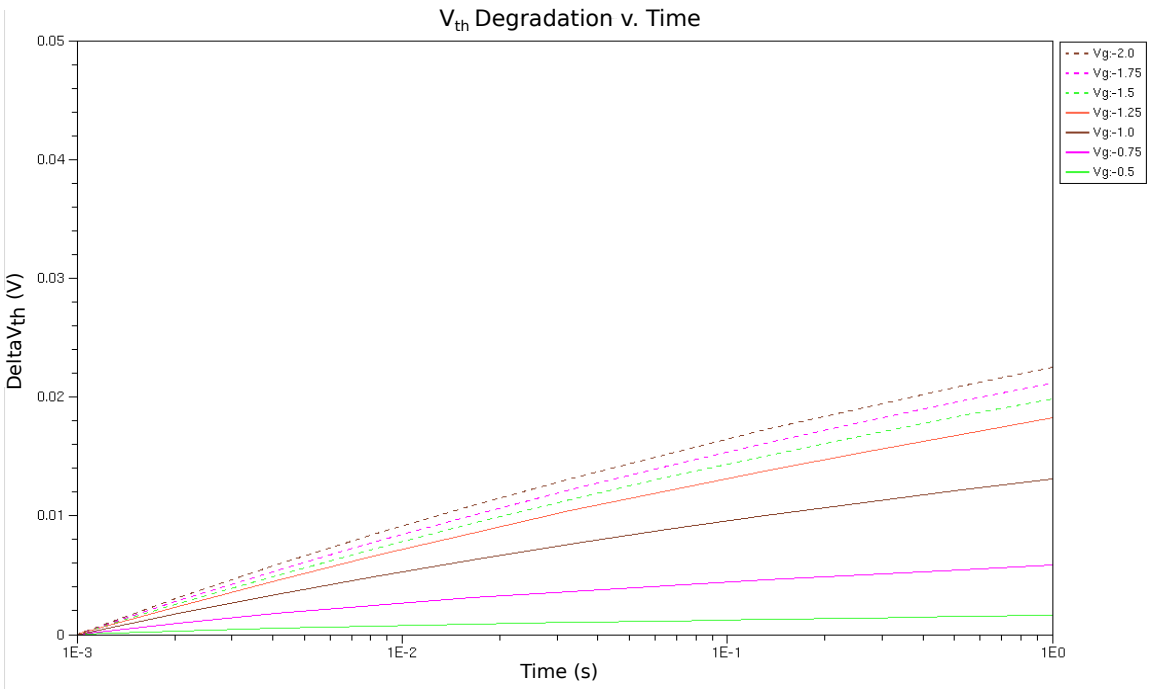


Figure 3.13: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 650$ K

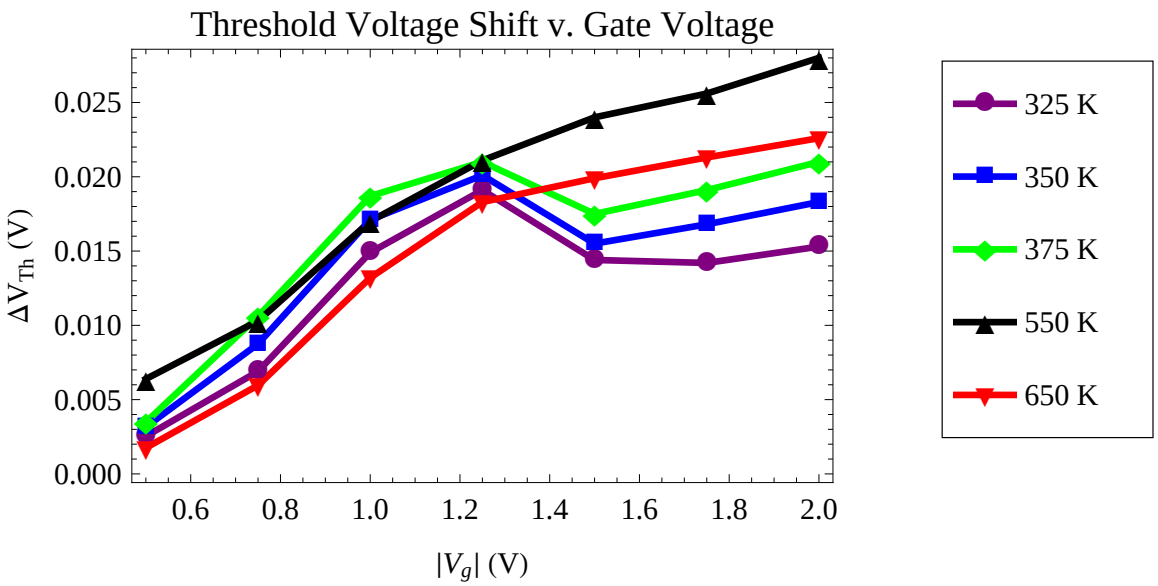


Figure 3.14: Variation of ΔV_{th} after 1 s symmetric stress for varying gate voltage and temperature

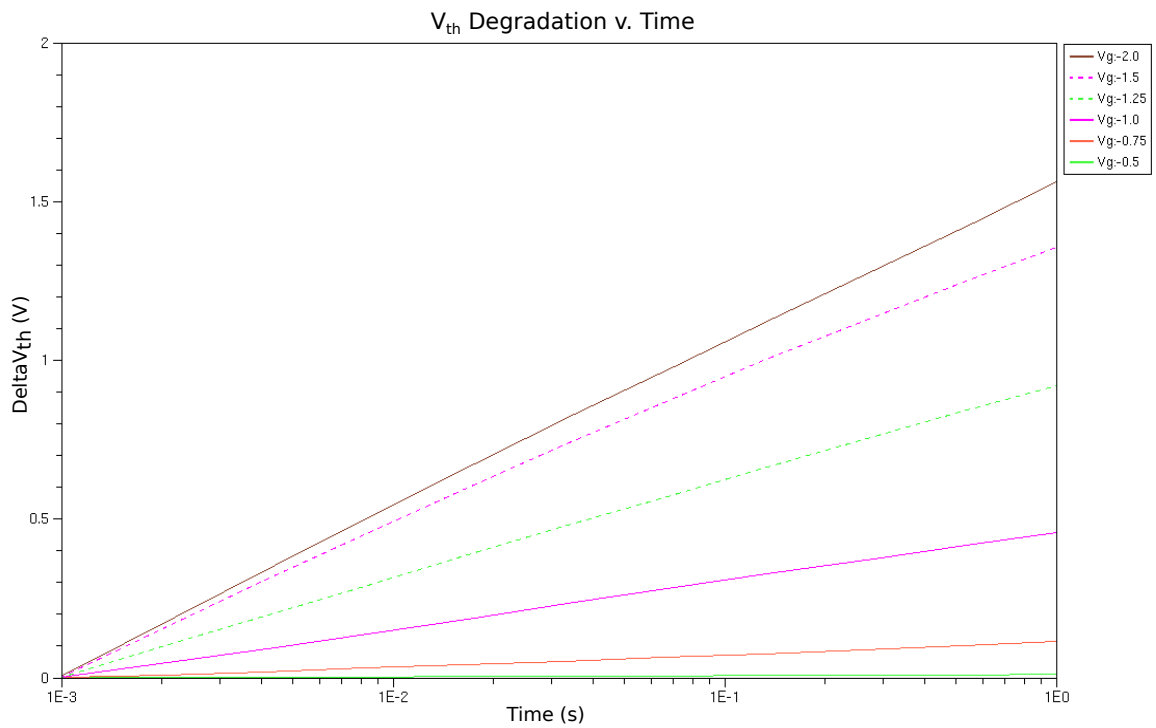


Figure 3.15: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during asymmetric stress ($V_{\text{g}} = V_d$), $T_{\text{ext}} = 350$ K

For completeness, the recovery in V_{th} following NBTI stress is shown in Figure 3.16, Figure 3.17, Figure 3.18, and Figure 3.19. The relaxation phase exhibits a longer recovery time compared to the degradation time during stressing, and the degradation no longer depends logarithmically on time. This asymmetry between stress and recovery originates from the slow recoverability of the P_b-centers (trap state 4), and the discrepancy in the probability for trap state 3 to relax back to trap state 1 compared to the probability for hole recapture (trap state 3 to state 2).

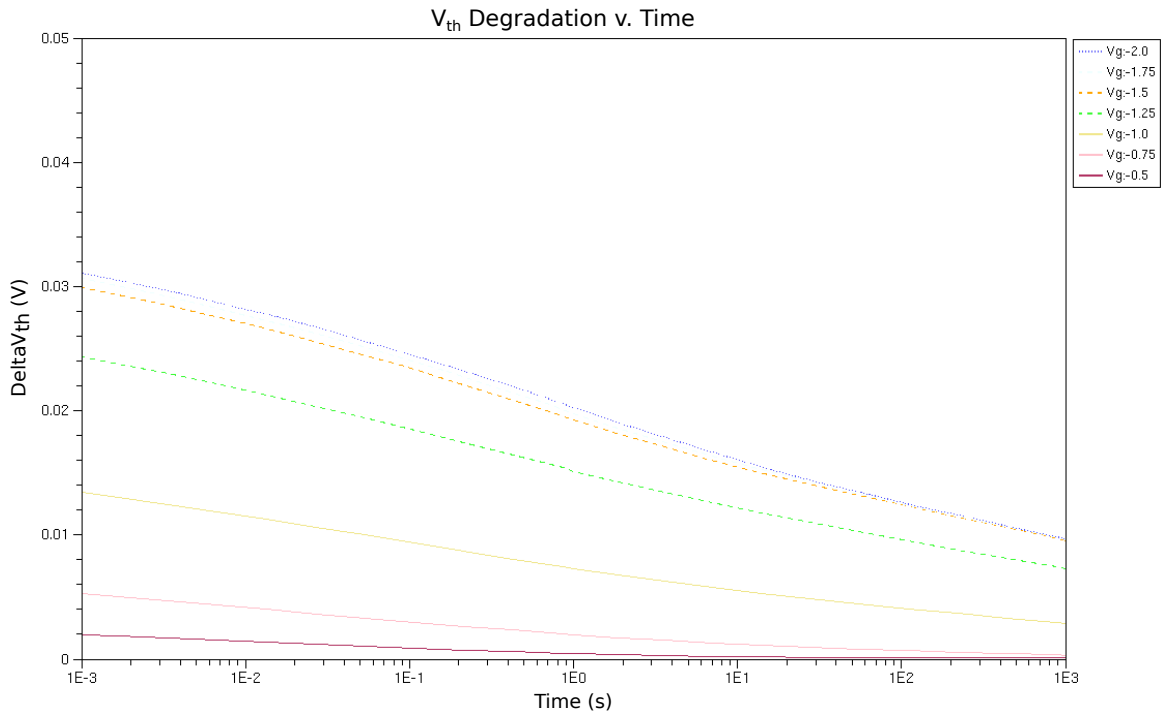


Figure 3.16: Recovery of V_{th} degradation following symmetric stress, $T_{\text{ext}} = 325$ K

Due to the complexity of competing mechanisms contributing to NBTI degradation, it is difficult to characterize the origin of broken monotonic behavior with respect to gate voltage exhibited during stressing at lower temperatures. However, we can gain some traction by considering the dynamics of the NBTI trap states 1-4 within this temperature limit. First, we examine the shift in $f_i(t)$ curves for fixed gate bias and varying temperature. This behavior is given in Figure 3.20 to Figure 3.31.

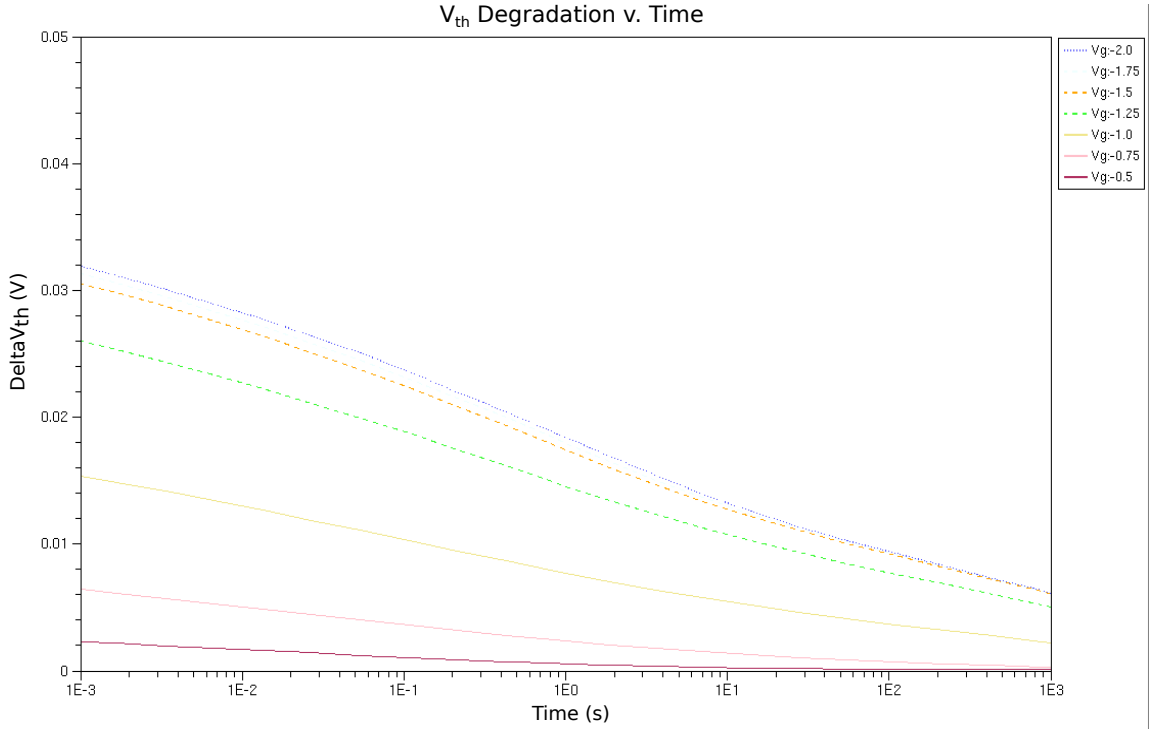


Figure 3.17: Recovery of V_{th} degradation following symmetric stress, $T_{\text{ext}} = 350$ K

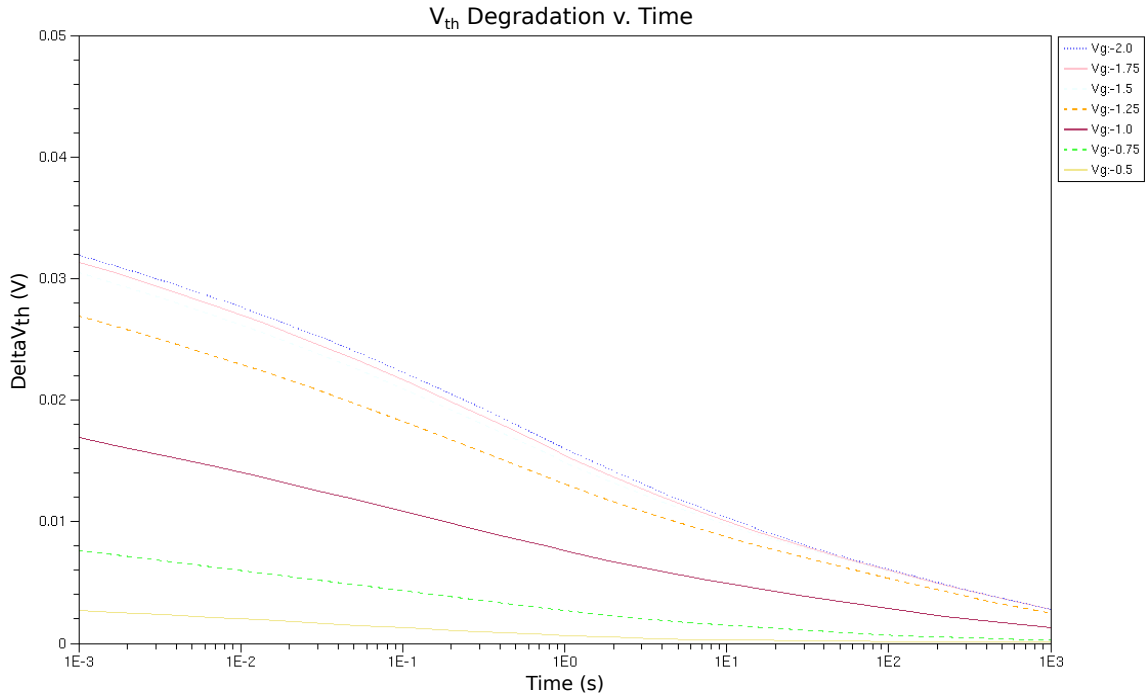


Figure 3.18: Recovery of V_{th} degradation following symmetric stress, $T_{\text{ext}} = 375$ K

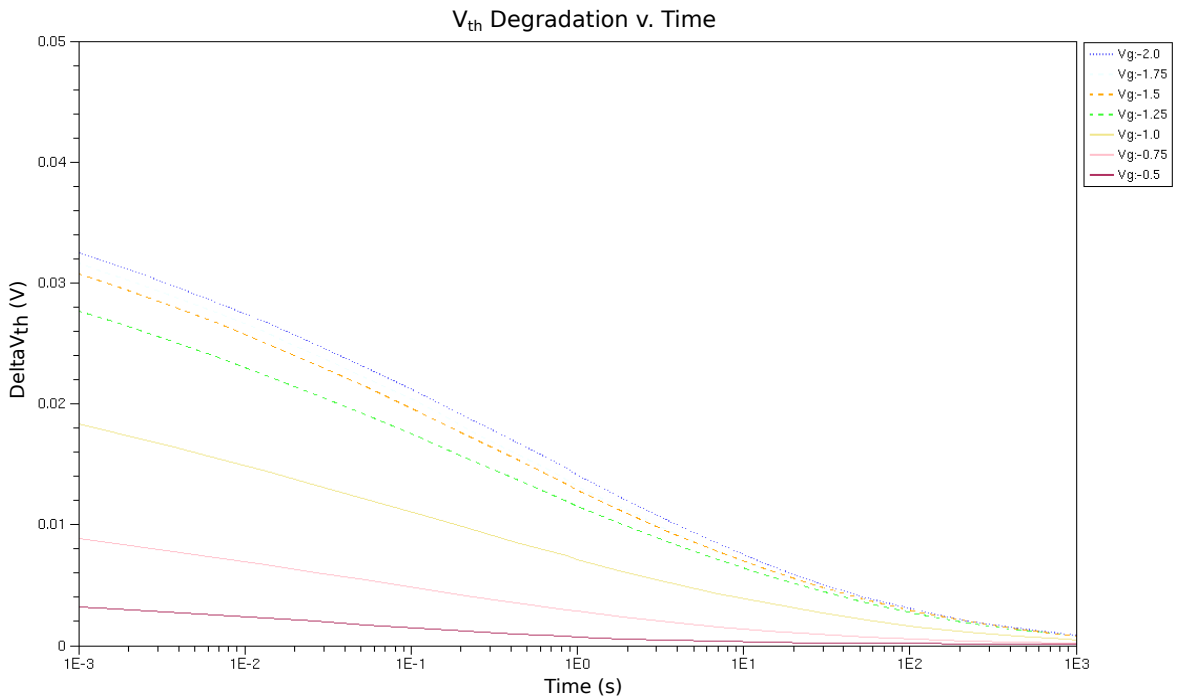


Figure 3.19: Recovery of V_{th} degradation following symmetric stress, $T_{\text{ext}} = 400$ K

It is important to note the respective time windows for the V_{th} degradation plots and the f_i trap state plots. While the simulation data for the trap state plots begins at $t = 10^{-15}$ s, the data for V_{th} degradation does not begin until $t = 10^{-3}$ s. Thus, the relevant range to examine the occupation of the trap states is $10^{-3} \leq t \leq 1$ s. Indeed, the simulation time step quickly approaches this scale after a couple iterations. The initial behavior of the trap states is indicative of noise in the discretized two-stage NBTI model, and the simulation exhibits a particular sensitivity to the initial time step.[9]

It is interesting to note that, within the appropriate time domain $10^{-3} \leq t \leq 1$ s, the slope and ordering of the trap state occupation curves f_1, f_3 and f_4 at different temperatures remain the same for increasing V_{g}. However, for trap state 2, we notice a markedly different behavior. At $V_{\text{g}} = -1.5$ V, the slope and ordering reverse. If

[9]Indeed, it was discovered retroactively that convergence often improved when the initial time step was increased, which is rather counterintuitive.

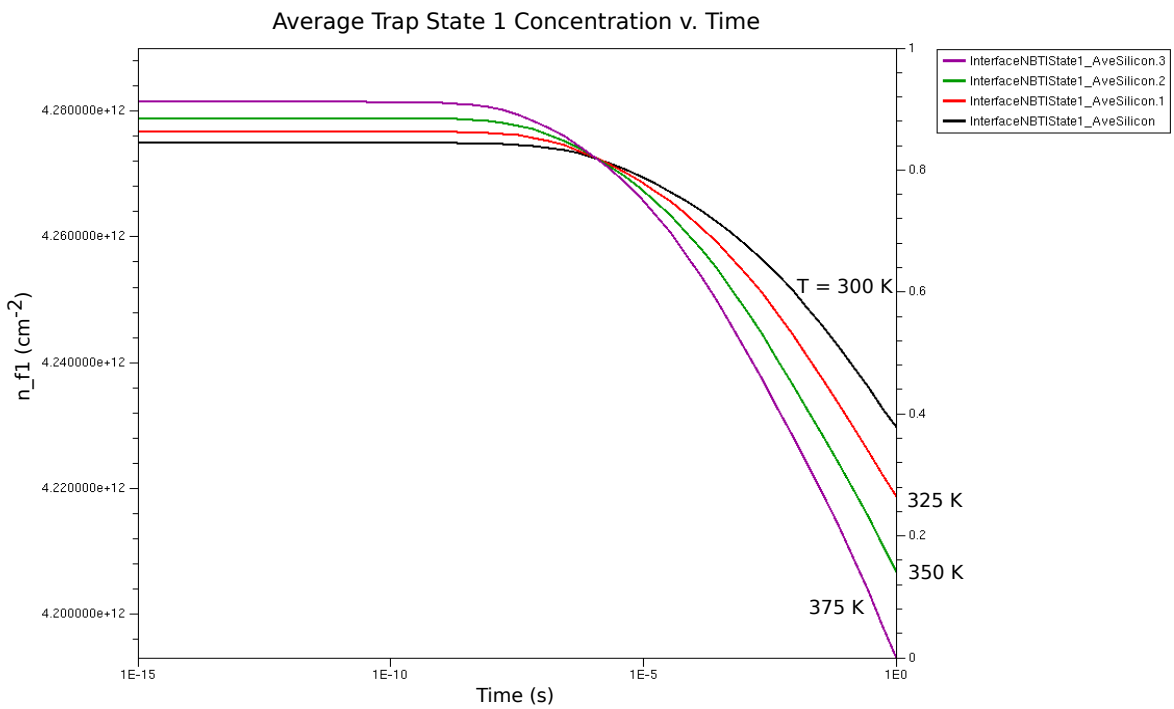


Figure 3.20: Temperature variation of $\langle f_1 \rangle$ during symmetric stress, $V_g = -0.5$ V

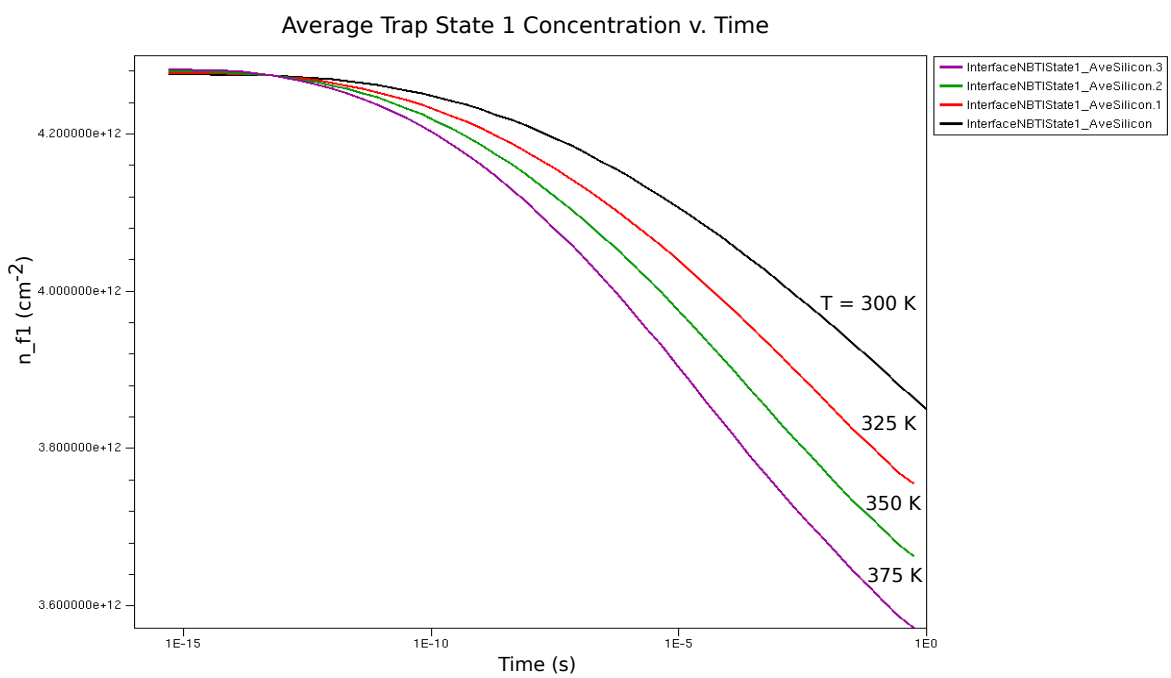


Figure 3.21: Temperature variation of $\langle f_1 \rangle$ during symmetric stress, $V_g = -1.0$ V

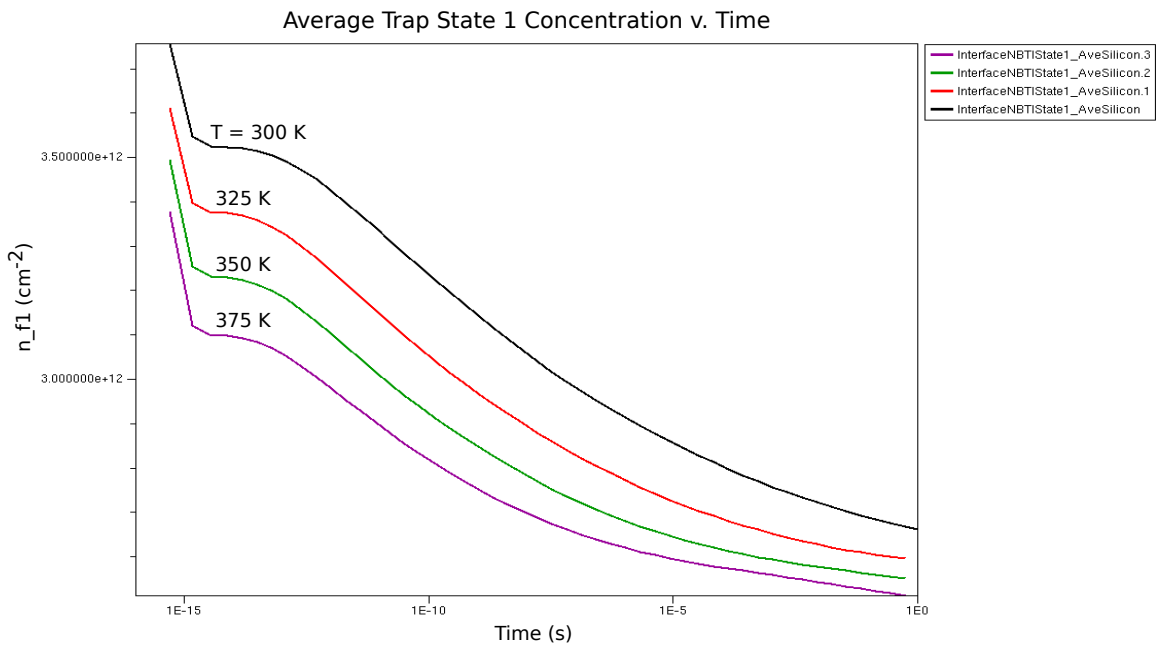


Figure 3.22: Temperature variation of $\langle f_1 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -1.5$ V

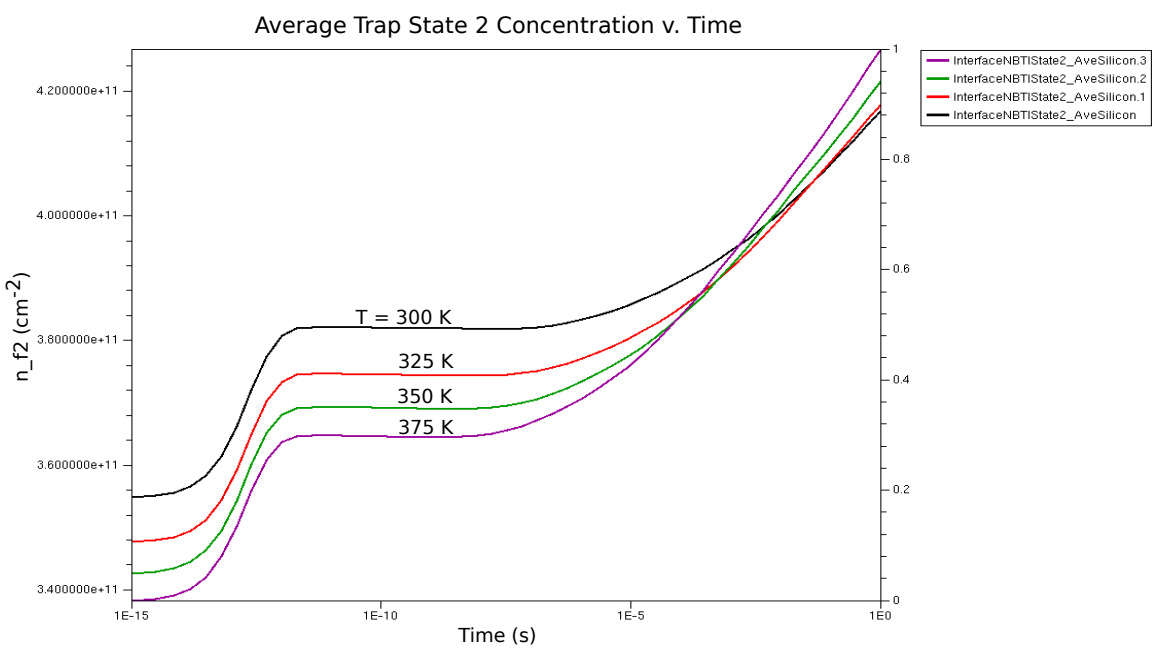


Figure 3.23: Temperature variation of $\langle f_2 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -0.5$ V

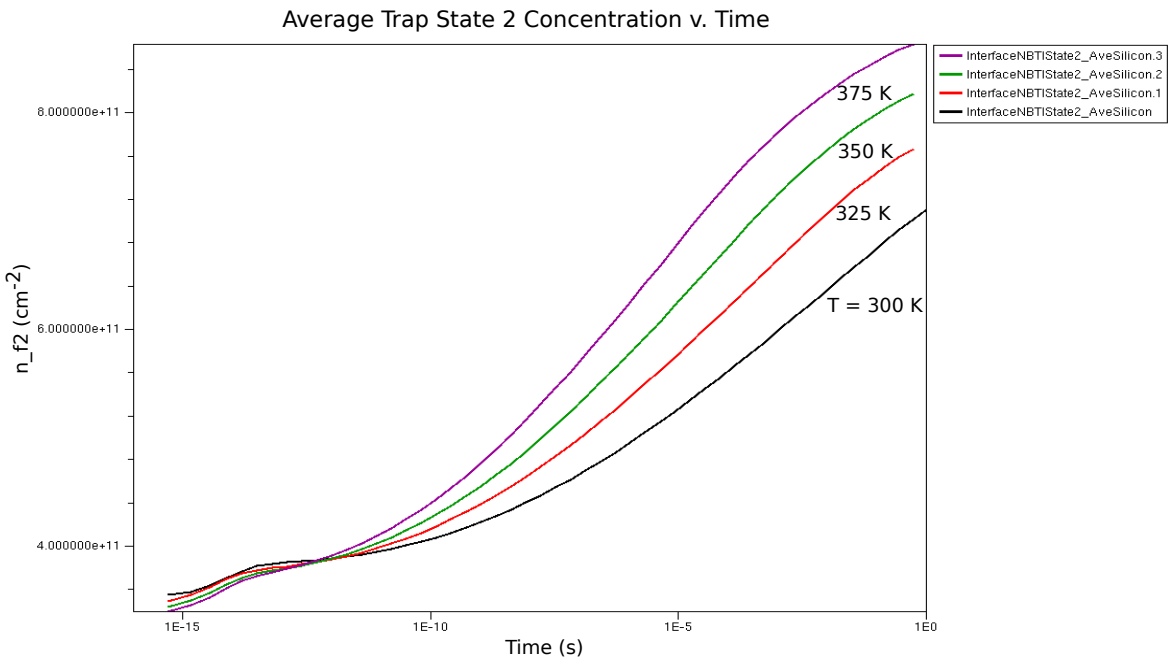


Figure 3.24: Temperature variation of $\langle f_2 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -1.0$ V

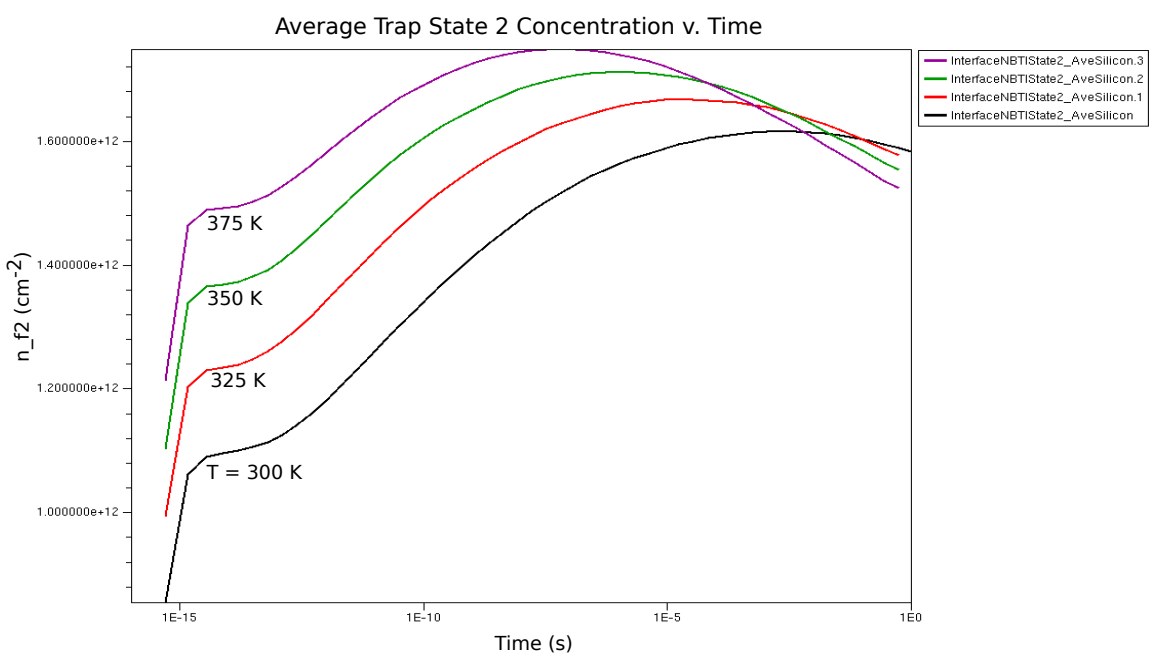


Figure 3.25: Temperature variation of $\langle f_2 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -1.5$ V

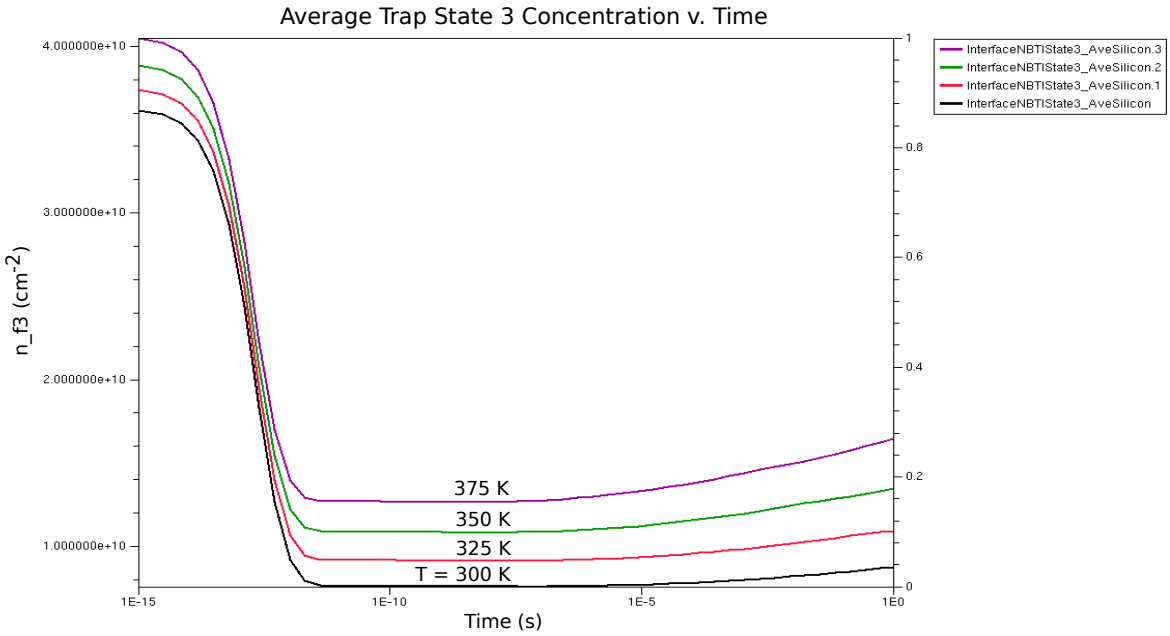


Figure 3.26: Temperature variation of $\langle f_3 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -0.5$ V

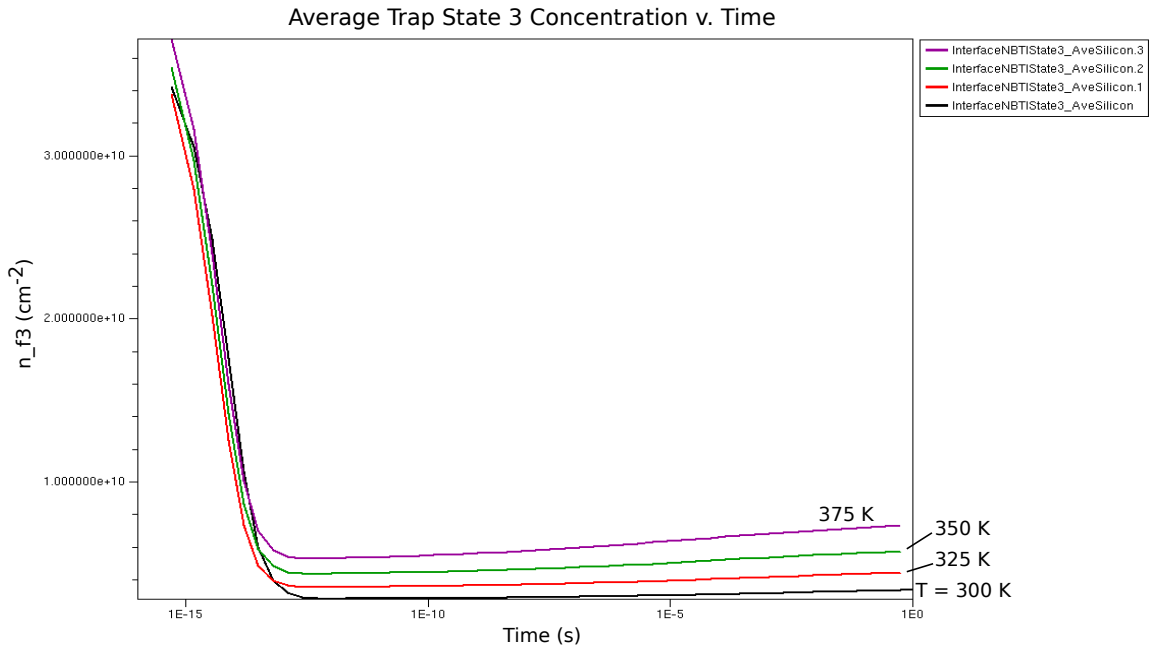


Figure 3.27: Temperature variation of $\langle f_3 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -1.0$ V

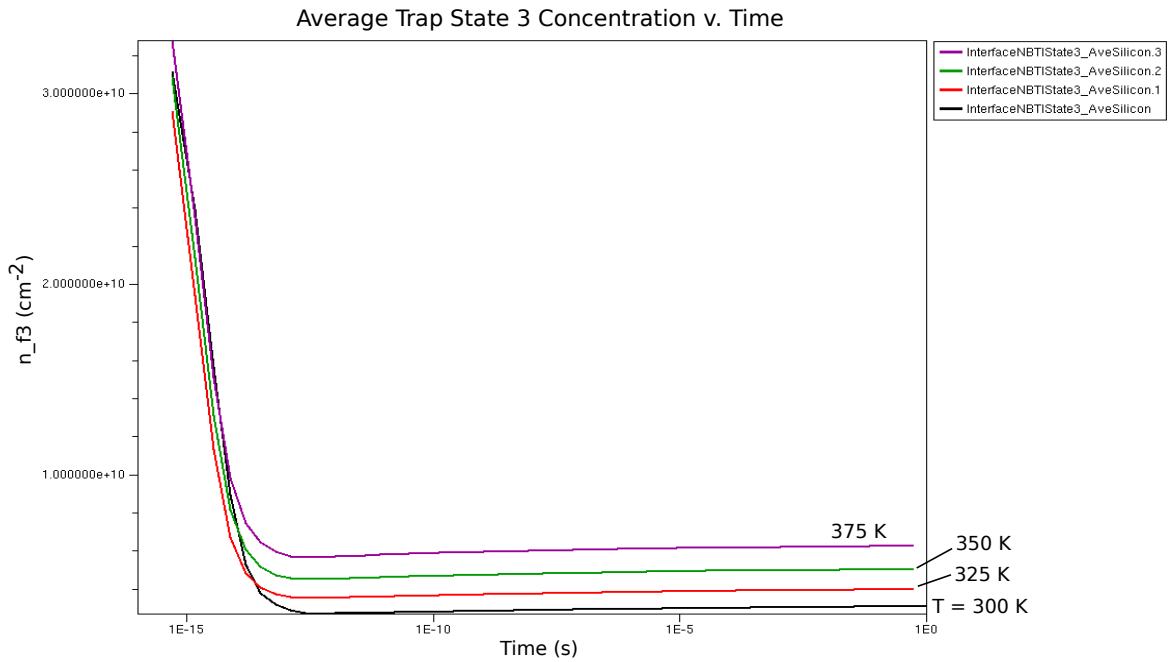


Figure 3.28: Temperature variation of $\langle f_3 \rangle$ during symmetric stress, $V_g = -1.5$ V

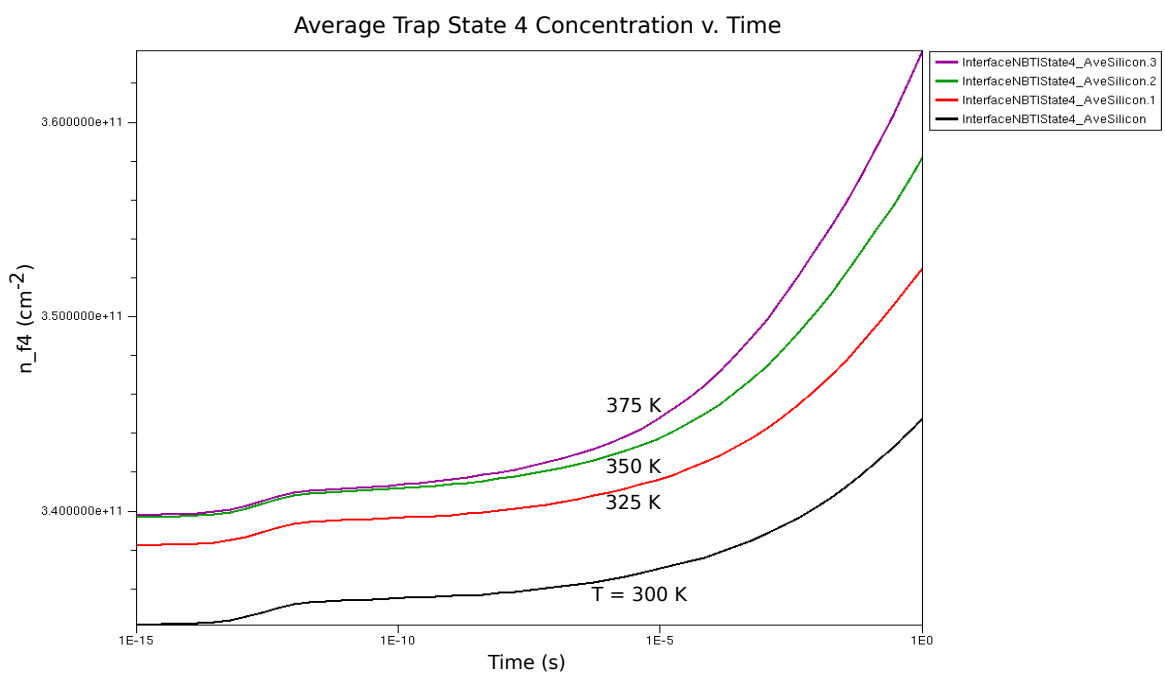


Figure 3.29: Temperature variation of $\langle f_4 \rangle$ during symmetric stress, $V_g = -0.5$ V

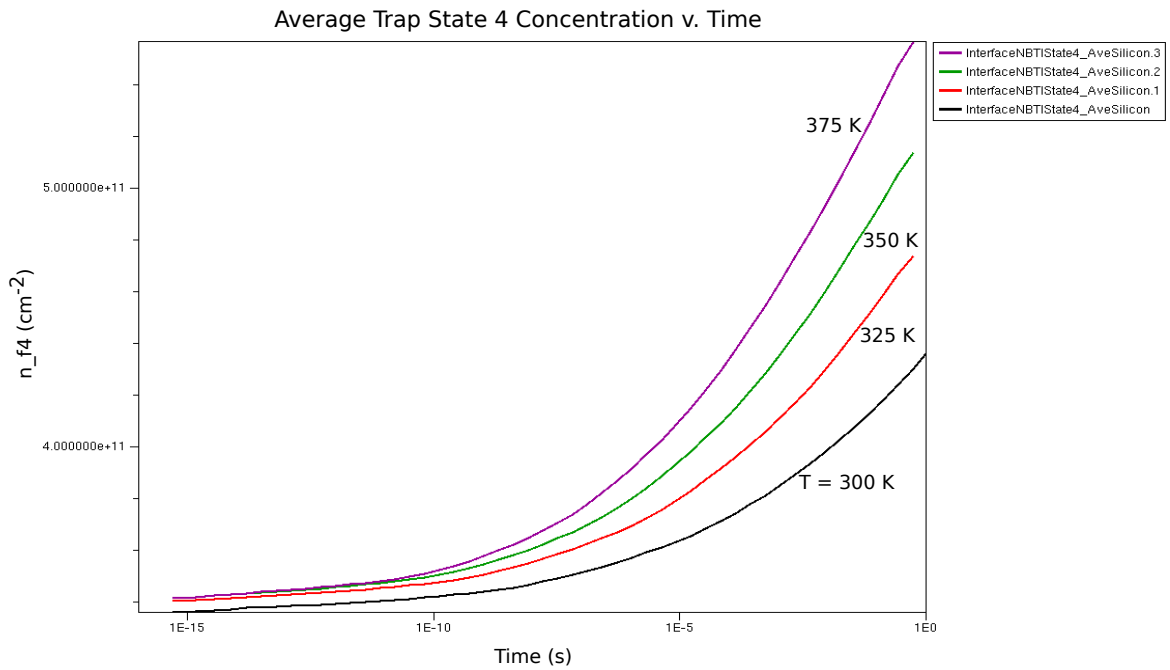


Figure 3.30: Temperature variation of $\langle f_4 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -1.0$ V

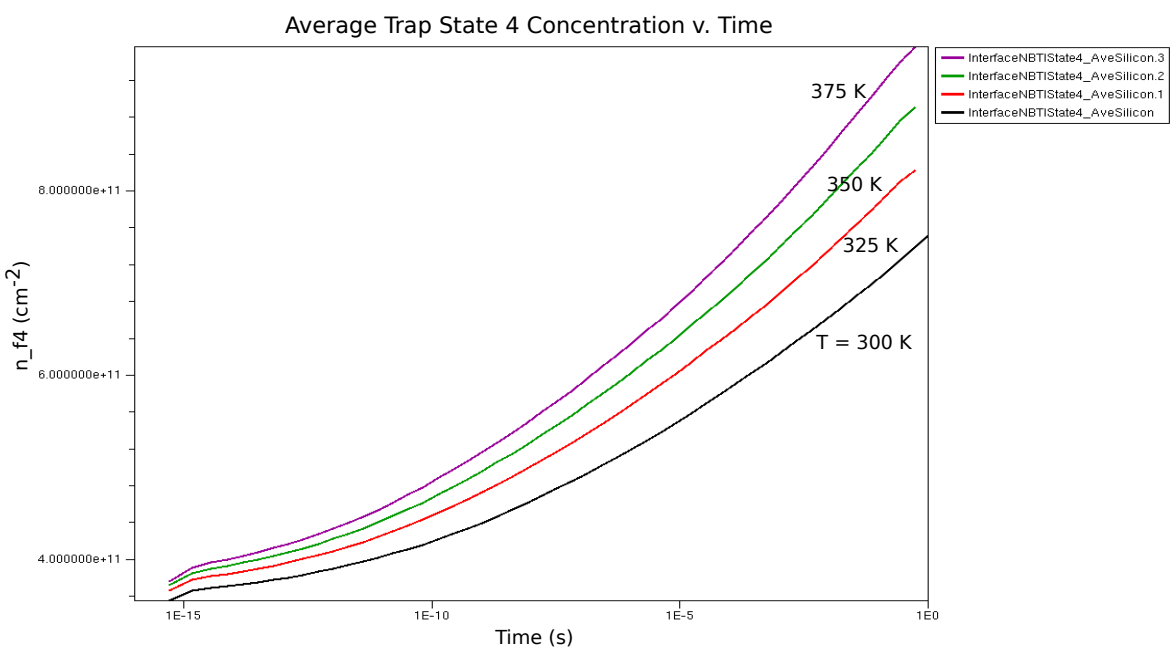


Figure 3.31: Temperature variation of $\langle f_4 \rangle$ during symmetric stress, $V_{\mathrm{g}} = -1.5$ V

we consider the transition pathways for state 2, then the number of holes in states 3 and 4 should be correlated to this behavior. Indeed, we find that while trap state 3 saturates around 5×10^{12} cm^{-2}, trap state 4 continues to increase in occupation number. That trap state 3 can recover to the precursor state 1, which is then available to transition to state 2, while state 4 has a low probability of recovery, accounts for the change in the transient behavior of state 2. We thus find that the generation of P_b-centers becomes the dominant transition pathway for state 2 at high gate bias, and for higher stress times we should expect to see the number of P_b-center defects exceed the number of E'_γ-center defects. We should, then, expect a different long-time behavior for NBTI degradation.

Nonetheless, if we examine the buildup of trapped charge in the oxide, we find that the amount of trapped charge Q increases as expected with increasing gate bias (Figure 3.32, Figure 3.33, and Figure 3.34). Furthermore, we find that these curves retain their ordering and slope for increasing V_g. If the behavior of trap state 2 accounted for the non-monotonic behavior of $\Delta V_{th}(t)$, then we would find a similar reordering or non-positive-definite behavior for $Q(t)$. That this is not the case implies that an external mechanism is responsible for the degradation in the V_{th}.

For completeness, we compare the behavior of trap states 1-4 and the total trapped charge in the low and high temperature limits for fixed ambient temperature and varying gate voltage (Figure 3.35 to Figure 3.44). We consider $T_{ext} = 350$ K for the low temperature behavior and $T_{ext} = 550$ K for the high temperature behavior. It is important to note that the time domain for the $T_{ext} = 550$ K plots begins at $t = 10^{-6}$ s. This change was made retroactively to improve numerical convergence, as mentioned above.

Again, we find the behavior of states 1, 3, and 4 consistent with what has been observed above. For trap state 2, we find that the slope of the curve becomes increasingly negative for higher V_g in both case, while for $T_{ext} = 550$ K we note the crossing

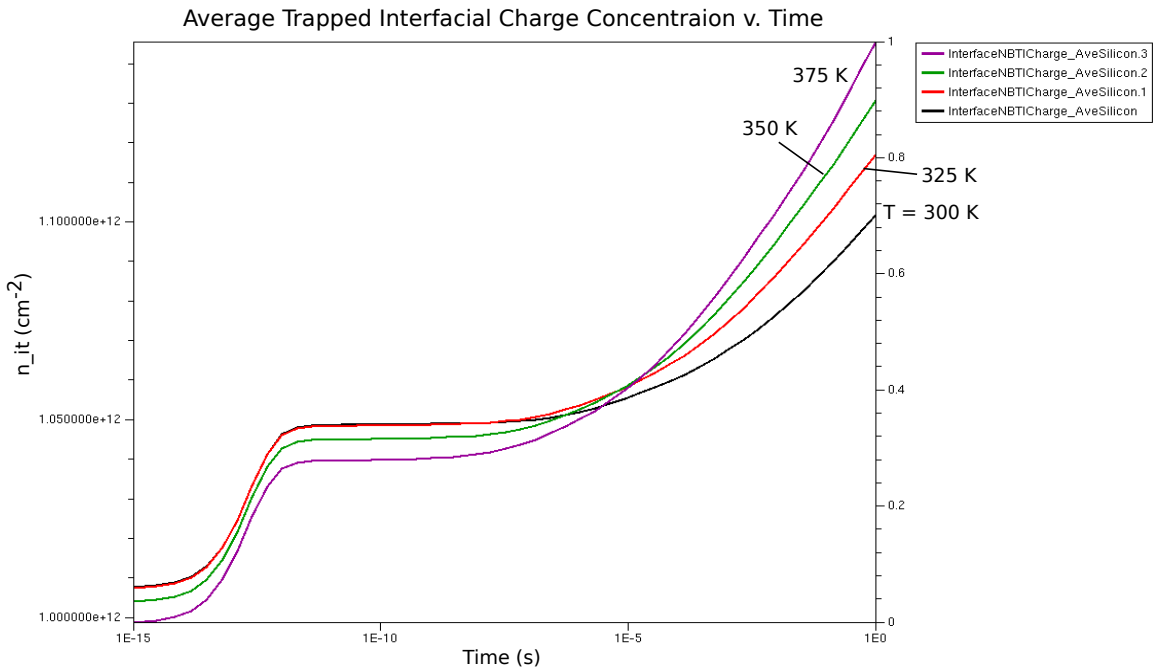


Figure 3.32: Temperature variation of Q during symmetric stress, $V_g = -0.5$ V

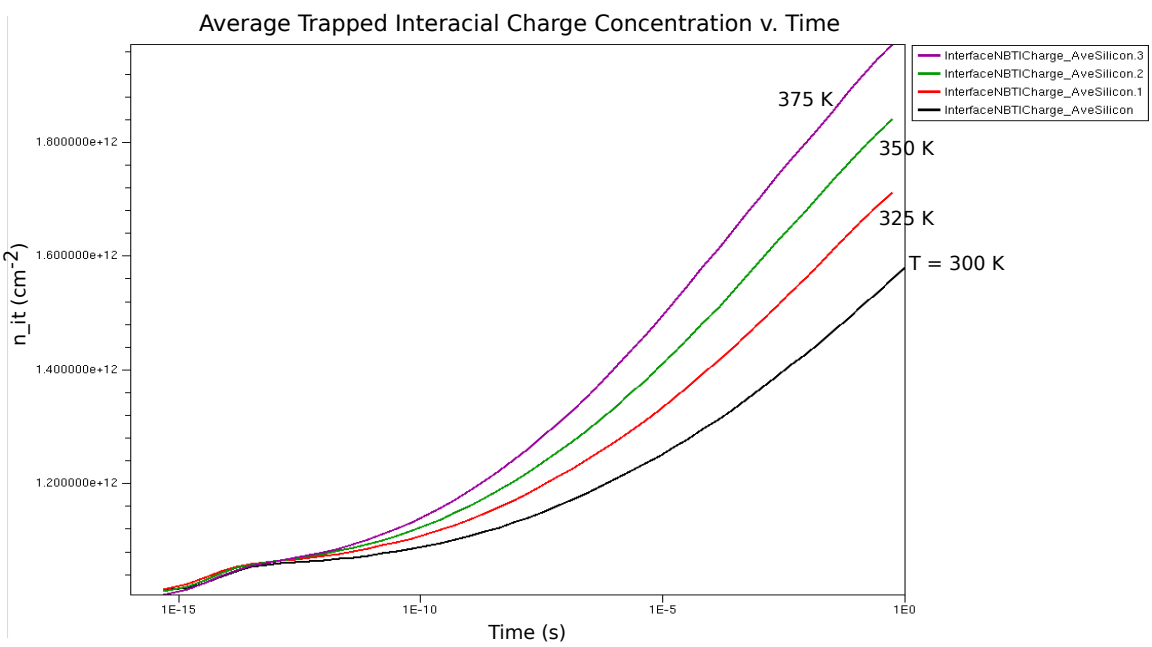


Figure 3.33: Temperature variation of Q during symmetric stress, $V_g = -1.0$ V

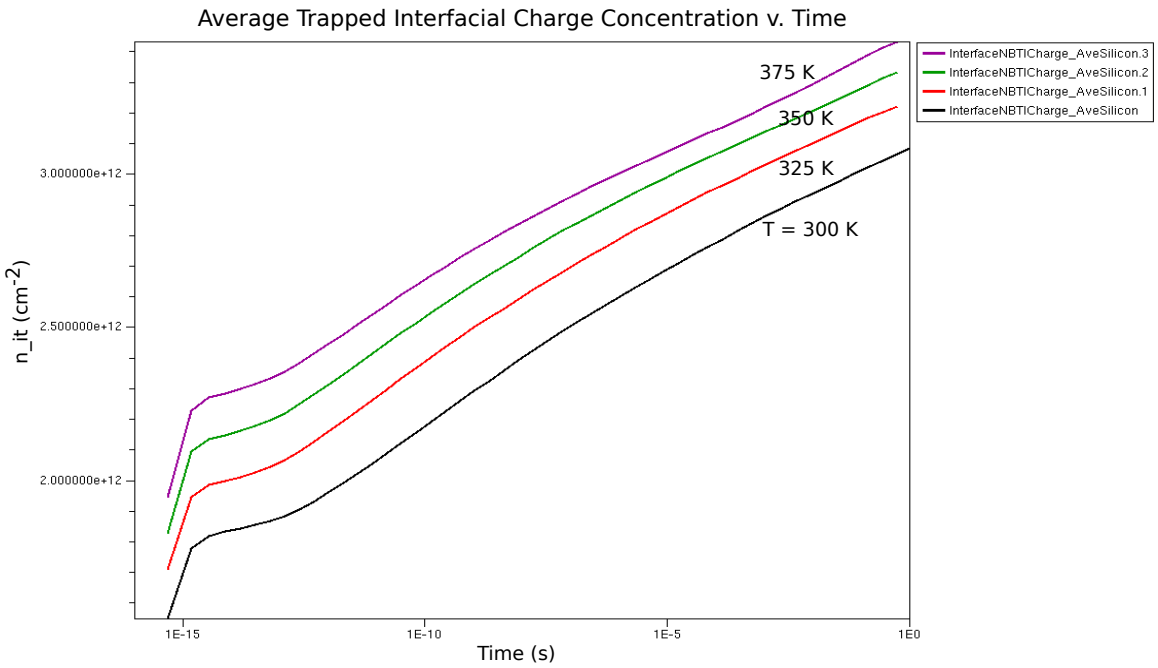


Figure 3.34: Temperature variation of Q during symmetric stress, $V_{\mathrm{g}} = -1.5$ V

between curves of higher V_{g}. The latter is indicative of a saturation in the transition rate from state 2 balanced with the availability of states 3 and 4. Nonetheless, the dynamics of the trap states cannot explain the non-monotonic behavior of ΔV_{th} at low temperatures and the restoration of monotonic behavior at high temperatures. Thus, we must consider other competing mechanisms.

Below we present spatial distributions for hole velocity, electric field, hole density, hole energy, hole current density, and lattice temperature for the channel region between the gates of the 2D FinFET device. These distributions represent the steady state solutions to the transport equations during the stress simulation, and we have found that these quantities reach this steady state quickly (within 2 to 3 time steps) during the transient part of the simulation. If we consider the local maximum and minimum values for these respective distributions, we find that there is steady increase in these values for every quantity except the hole velocity. Rather, we find that the maximum of the hole velocity for $V_{\mathrm{g}} = -1.5$ V is less than the hole velocity

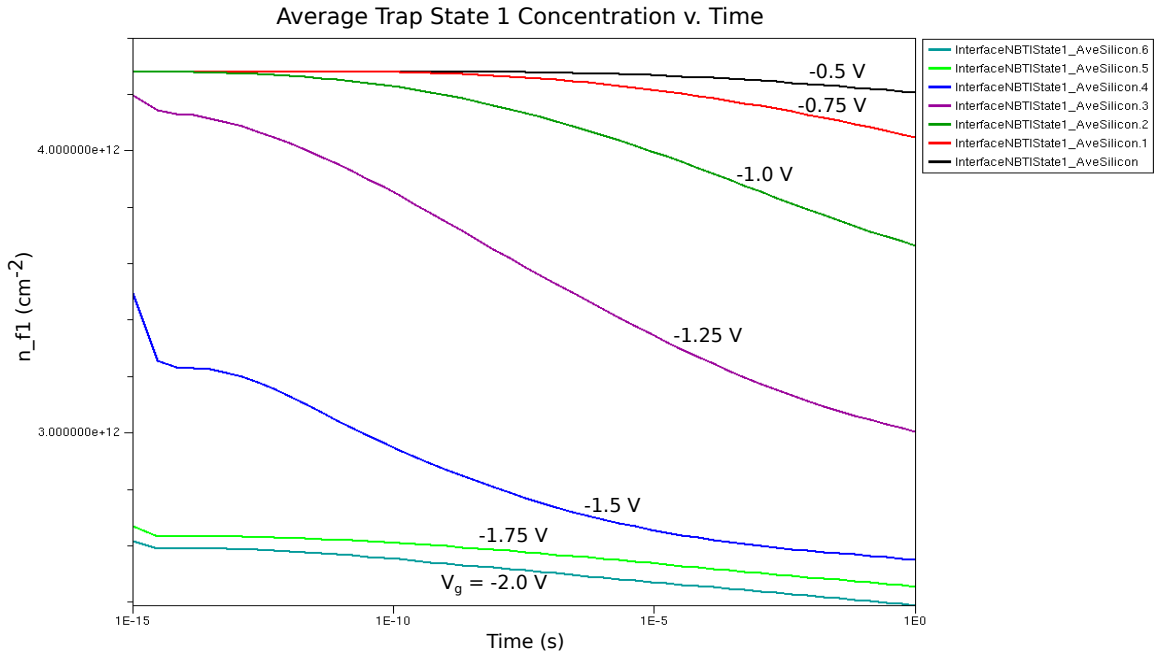


Figure 3.35: Gate voltage variation of $\langle f_1 \rangle$ during symmetric stress, $T_{\text{ext}} = 350$ K

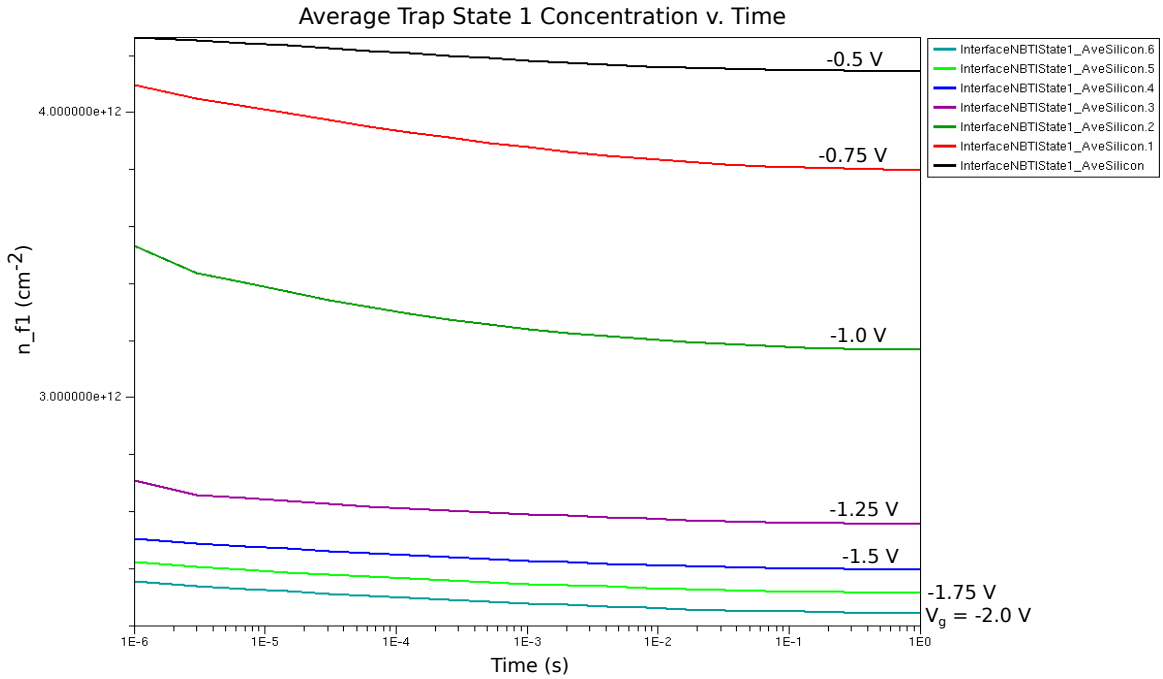


Figure 3.36: Gate voltage variation of $\langle f_1 \rangle$ during symmetric stress, $T_{\text{ext}} = 550$ K

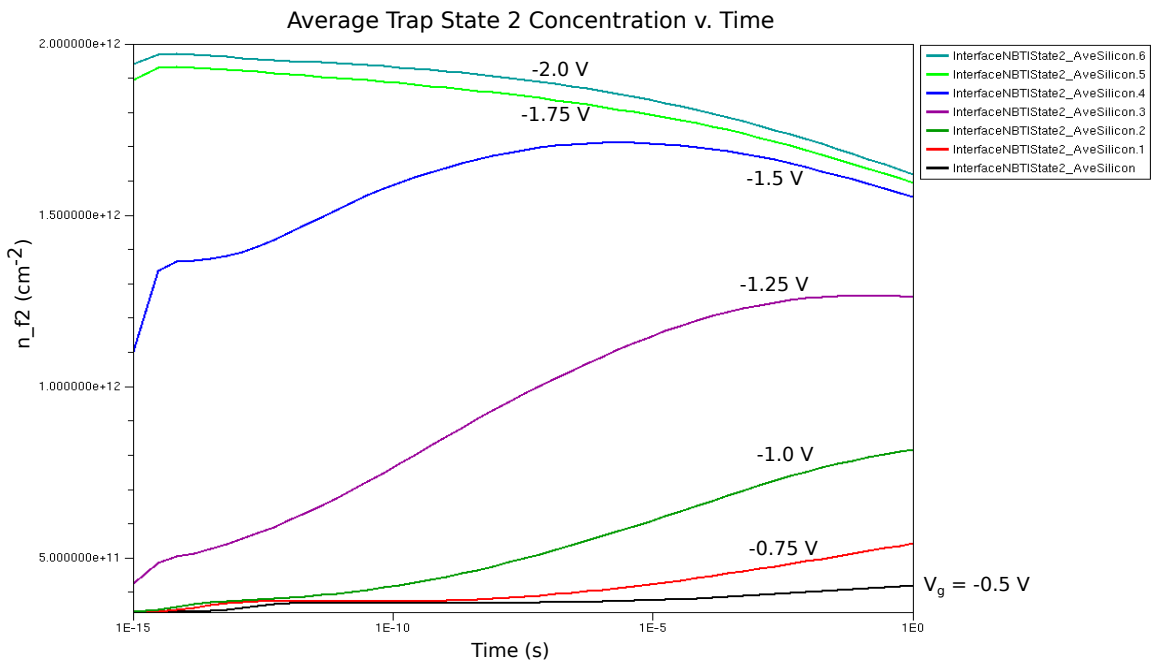


Figure 3.37: Gate voltage variation of $\langle f_2 \rangle$ during symmetric stress, $T_{\text{ext}} = 350$ K

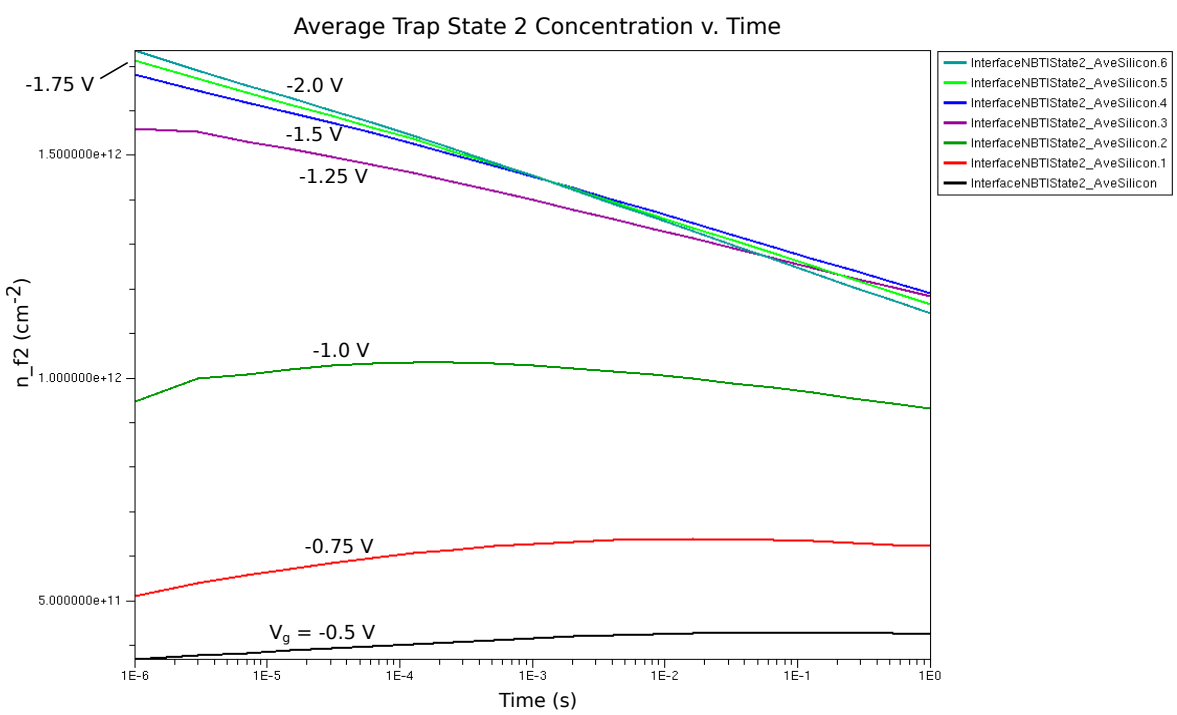


Figure 3.38: Gate voltage variation of $\langle f_2 \rangle$ during symmetric stress, $T_{\text{ext}} = 550$ K

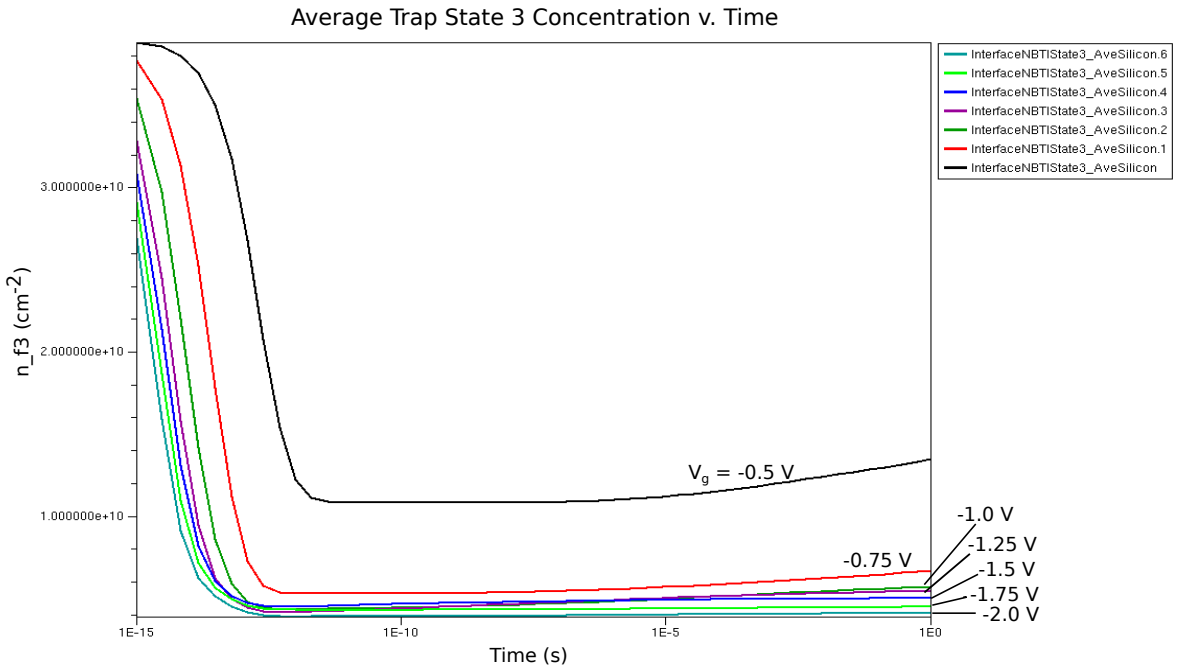


Figure 3.39: Gate voltage variation of $\langle f_3 \rangle$ during symmetric stress, $T_{\text{ext}} = 350$ K

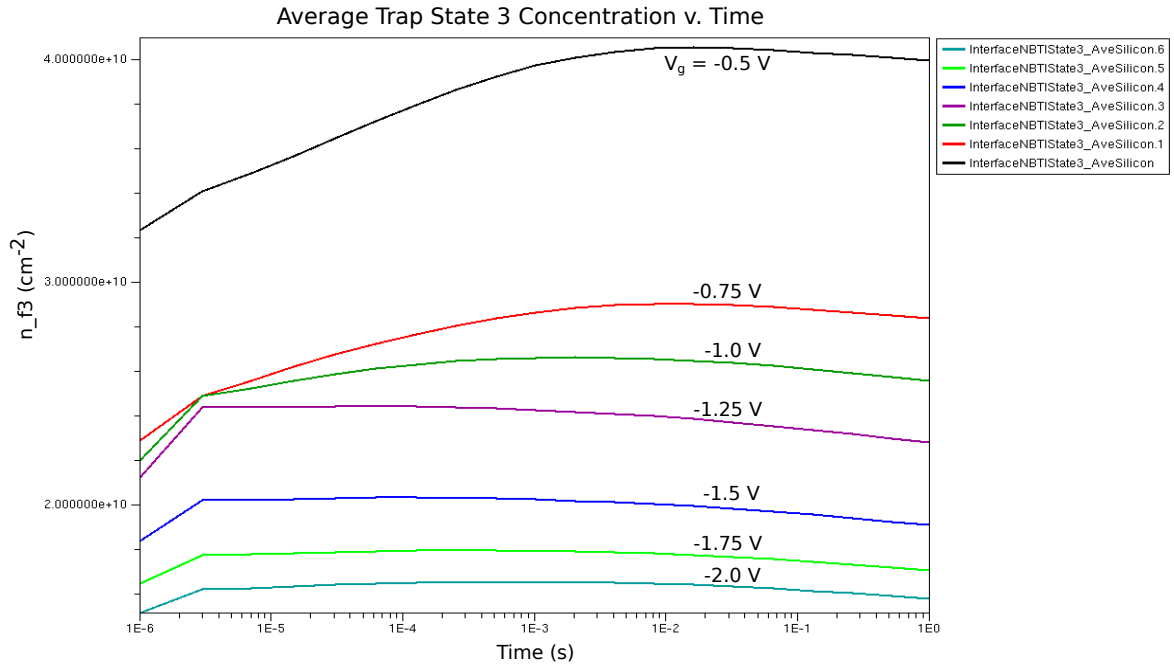


Figure 3.40: Gate voltage variation of $\langle f_3 \rangle$ during symmetric stress, $T_{\text{ext}} = 550$ K

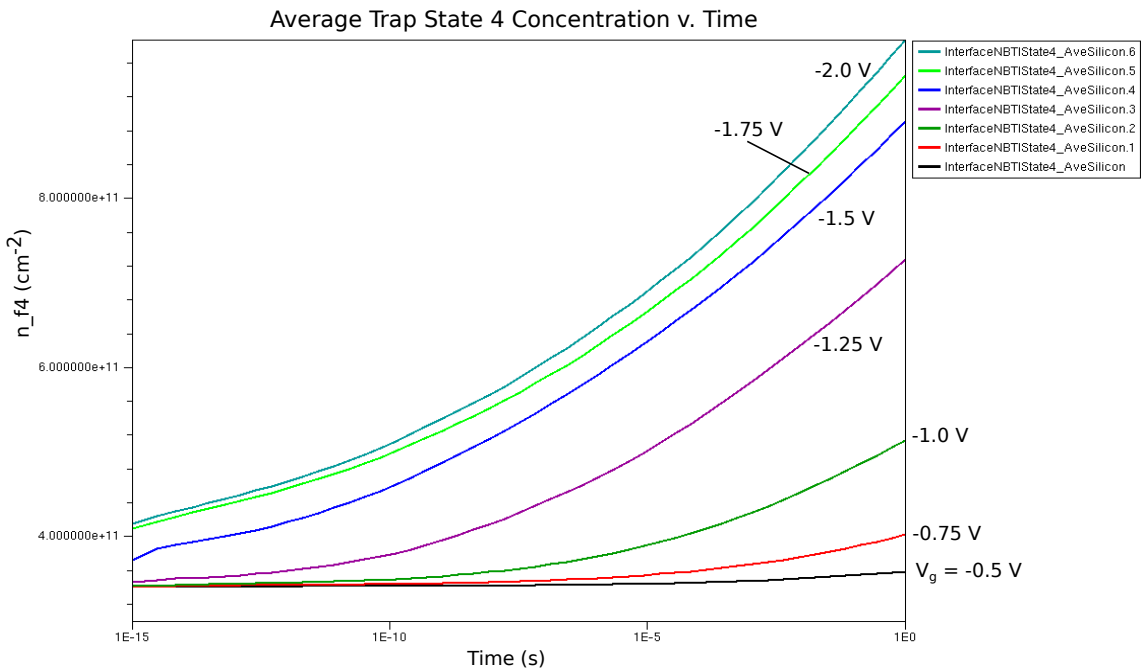


Figure 3.41: Gate voltage variation of $\langle f_4 \rangle$ during symmetric stress, $T_{\text{ext}} = 350$ K

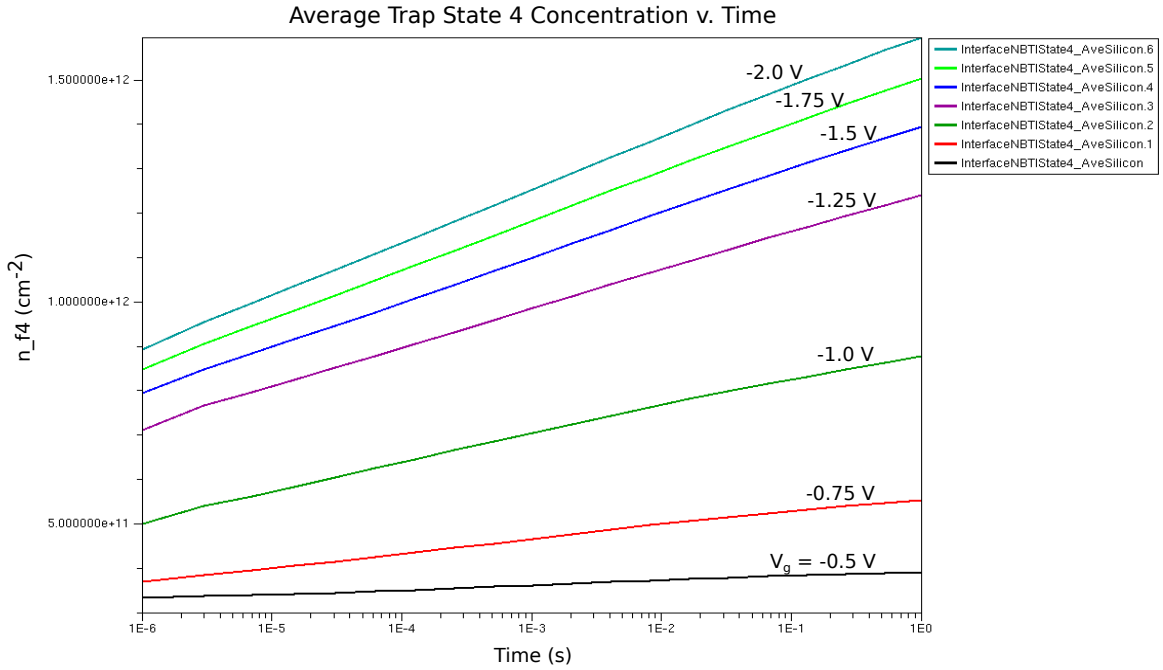


Figure 3.42: Gate voltage variation of $\langle f_4 \rangle$ during symmetric stress, $T_{\text{ext}} = 550$ K

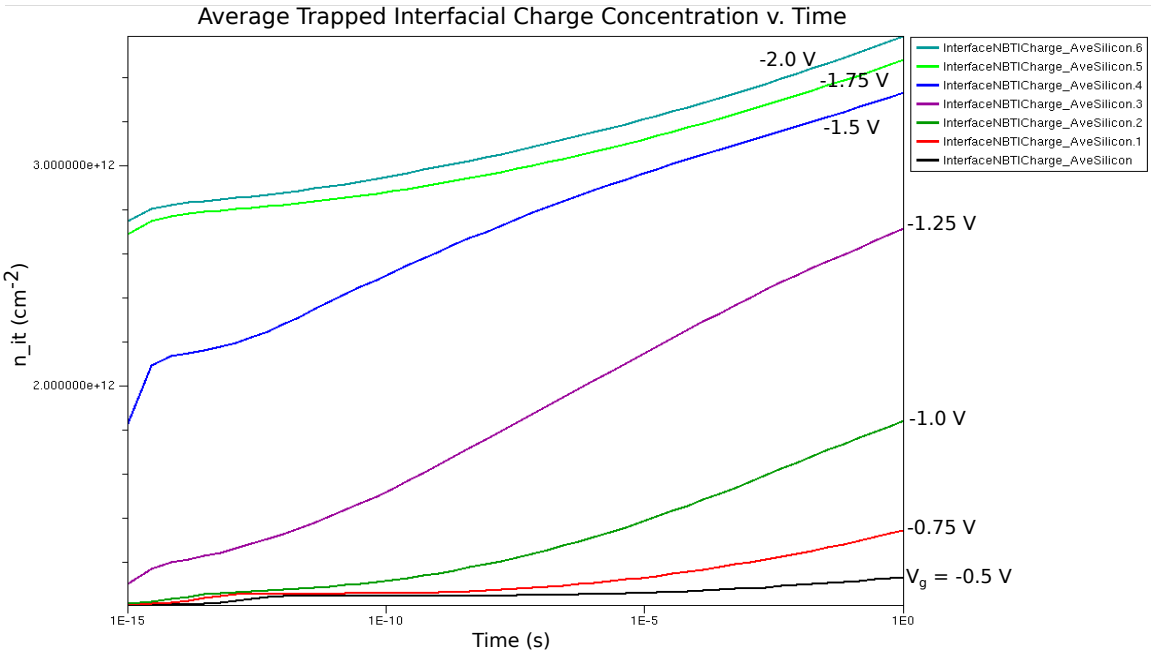


Figure 3.43: Gate voltage variation of Q during symmetric stress, $T_{\text{ext}} = 350$ K

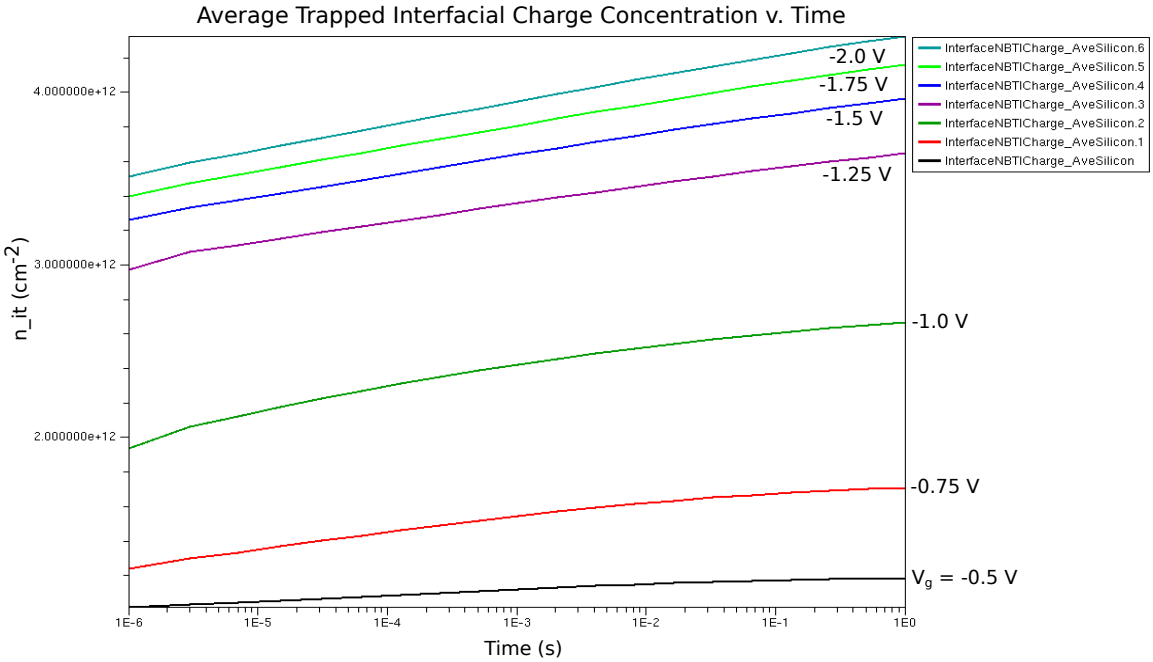
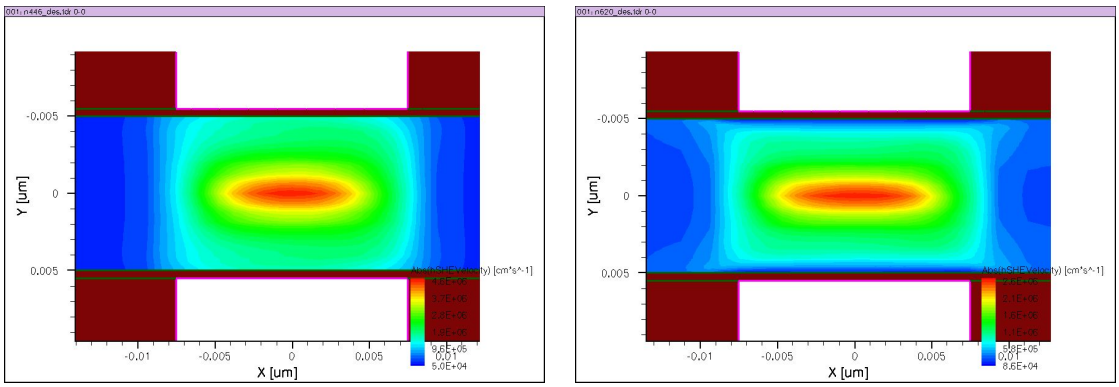


Figure 3.44: Gate voltage variation of Q during symmetric stress, $T_{\text{ext}} = 550$ K

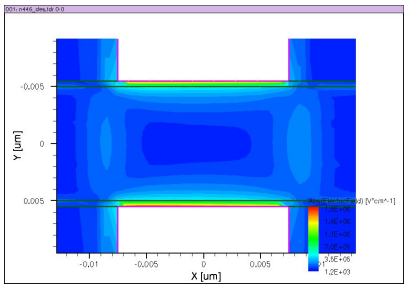
for $V_g = -0.5$ V, while the range for this quantity becomes narrower (Figure 3.45).



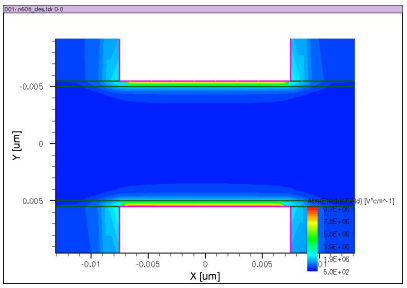
(a) $V_g = -0.5$ V, $T_{ext} = 350$ K (b) $V_g = -1.5$ V, $T_{ext} = 350$ K

Figure 3.45: Hole velocity $|\boldsymbol{v}_h|$ under symmetric stress

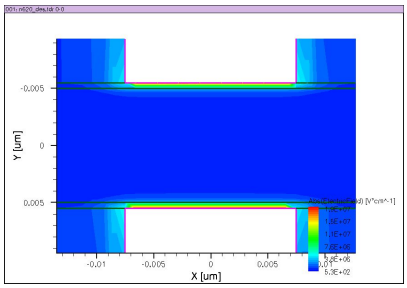
We note some general features of the electric field, hole density, hole energy, and hole current density distributions. In Figure 3.46, we find that the electric field maxima become tightly localized near the gate and extend into the oxynitride spacers as an effect of the corners. In Figure 3.47, the hole density exhibits similar qualitative behavior, as more holes accumulate beneath the gates in response to the electric field, leaving behind a depletion region in the center of the channel. The hole energy (Figure 3.48), on the other hand, exhibits a curious resonant behavior across the channel that is more pronounced for higher V_g, and this is likely due to the mathematical structure of the SHE approximation employed in the device transport equations. Finally, we notice that the hole current (Figure 3.49) is also highly localized beneath the gates following the hole distribution, while the local maximum at the center of the channel corresponds to a low vertical electric field and indicates the possibility for quasi-ballistic transport through the center of the device.



(a) $V_{\mathrm{g}} = -0.5$ V, $T_{\mathrm{ext}} = 350$ K

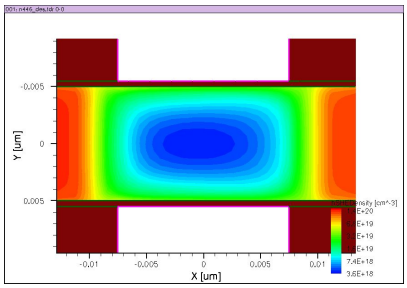


(b) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 350$ K

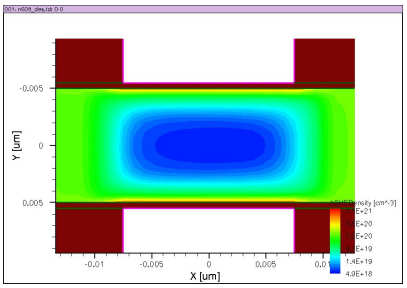


(c) $V_{\mathrm{g}} = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

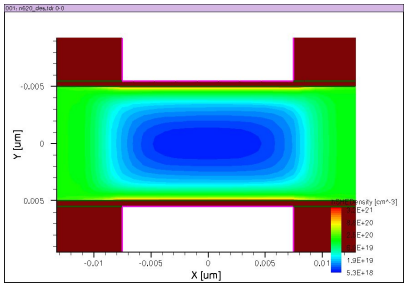
Figure 3.46: Electric field $|\boldsymbol{E}|$ under symmetric stress



(a) $V_{\mathrm{g}} = -0.5$ V, $T_{\mathrm{ext}} = 350$ K

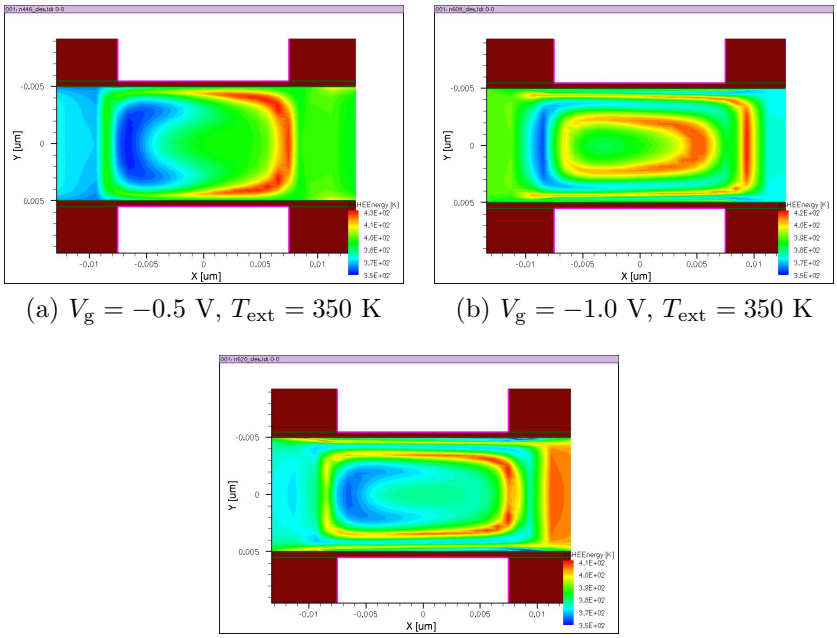


(b) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 350$ K



(c) $V_{\mathrm{g}} = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

Figure 3.47: Hole density n_{h} under symmetric stress

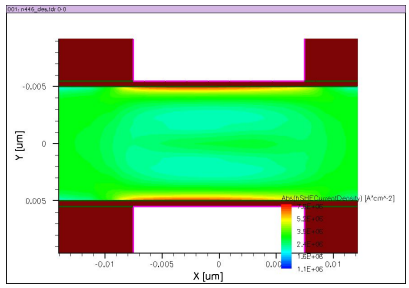


(a) $V_{\mathrm{g}} = -0.5$ V, $T_{\mathrm{ext}} = 350$ K (b) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 350$ K

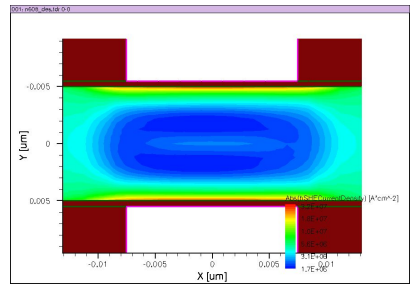
(c) $V_{\mathrm{g}} = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

Figure 3.48: Hole energy ϵ_{h} under symmetric stress

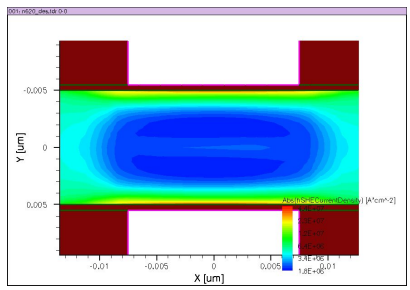
For symmetric stress configurations, we find that the lattice self-heating is not significant, as the device maintains an average temperature close to the boundary condition temperature with differences in the maximum and minimum temperatures on the order of 0.1 K. This is, in large part, due to the Dirichlet thermal boundary conditions imposed on the simulation. Indeed, we expect the formation of a hot spot in the channel to be insignificant due to the low bias applied to the drain. Figure 3.50 shows the spatial distribution of lattice temperature for $T_{\mathrm{ext}} = 350$ K. Qualitatively, the hot spot becomes more localized for higher V_{g}, but again the temperature rise is very little.



(a) $V_{\mathrm{g}} = -0.5$ V, $T_{\mathrm{ext}} = 350$ K



(b) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 350$ K



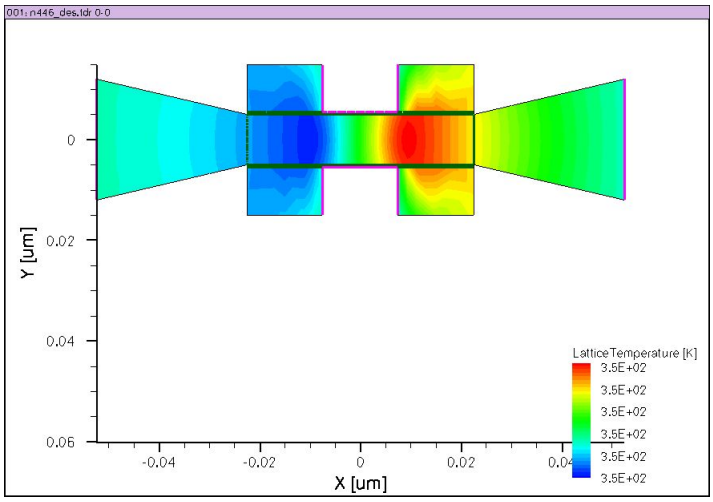
(c) $V_{\mathrm{g}} = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

Figure 3.49: Hole current density $|\boldsymbol{J}_{\mathrm{h}}|$ under symmetric stress

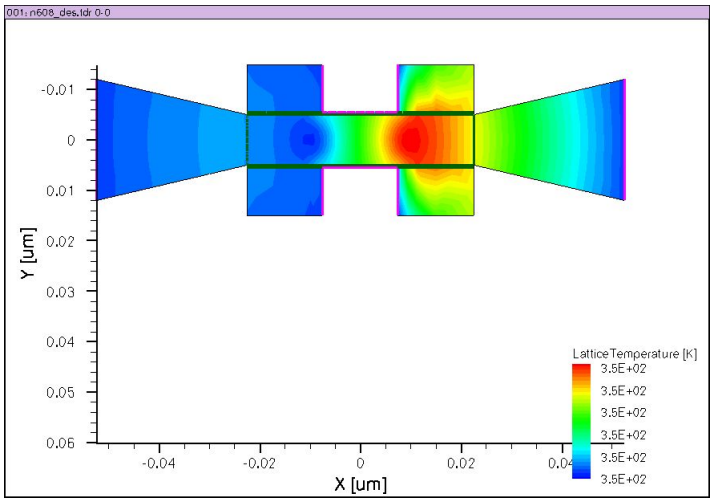
3.1.2 Asymmetric Stress Configuration

Under asymmetric stress conditions ($V_d = V_{\mathrm{g}}$), we observe a strikingly different behavior in the trap states for increasing V_{g}. We have kept $T_{\mathrm{ext}} = 350$ K in order to compare to the symmetric stress data. Plots for the dynamics of the occupation of the trap states and the total trapped charge are provided in Figure 3.51, Figure 3.52, Figure 3.53, Figure 3.54, and Figure 3.55. Note that the simulation failed to converge for $V_{\mathrm{g}} = -1.75$ V, hence its absence in the plots. In this asymmetric configuration, we find a reordering of the $f_i(t)$ curves for different V_{g}, and a corresponding behavior for $Q(t)$. This interesting dynamic is likely driven by the the structure of the electric field, the temperature gradient, and the deviation of the hole temperature away from the lattice temperature, as we discuss below.

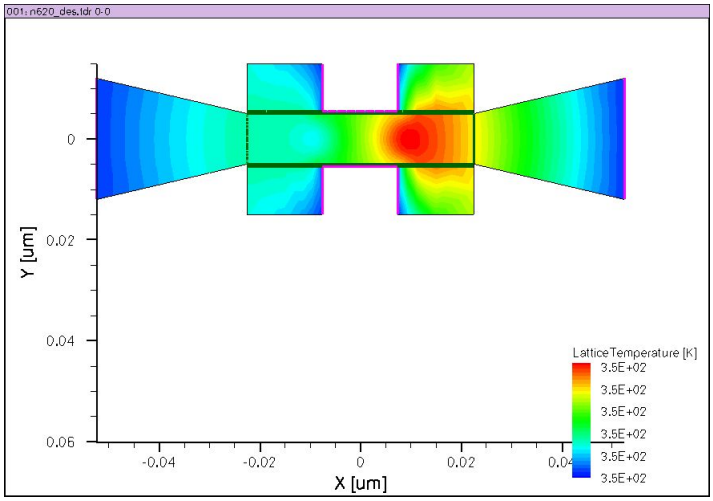
Figure 3.56 shows spatial distributions for the hole velocity for different V_{g}. These distributions are structurally different than in the symmetric case, as the holes in the



(a) $V_{\mathrm{g}} = -0.5$ V, $T_{\mathrm{ext}} = 350$ K



(b) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 350$ K



(c) $V_{\mathrm{g}} = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

Figure 3.50: Lattice temperature T_{L} under symmetric stress

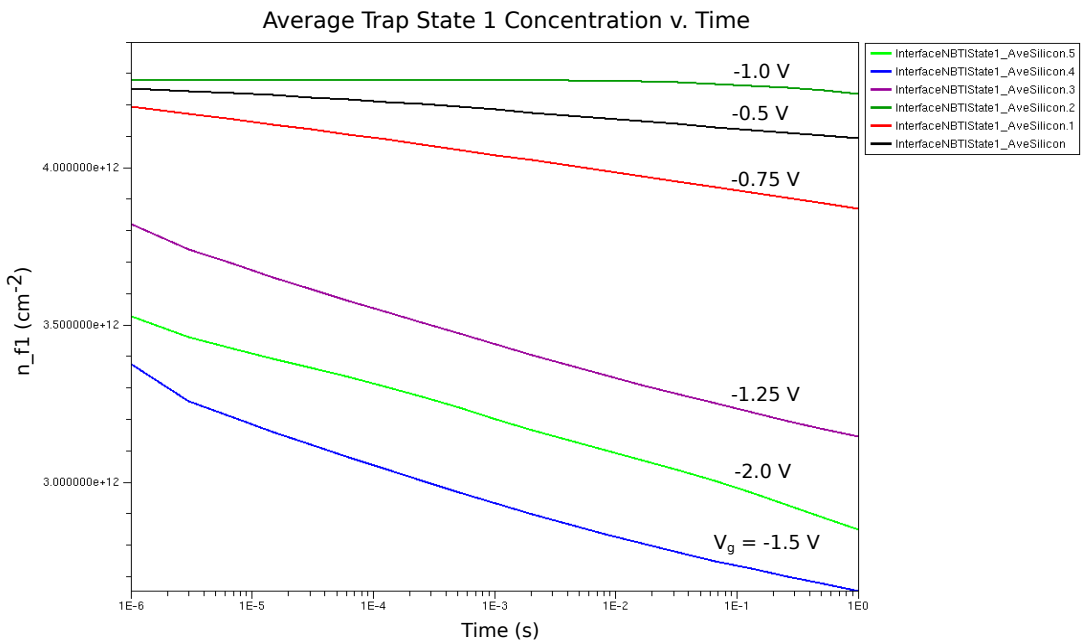


Figure 3.51: Gate voltage variation of $\langle f_1 \rangle$ during asymmetric stress, $T_{\text{ext}} = 350$ K

channel experience greater acceleration towards the drain due to the drain bias. The fan-out in the structure is due to the fringing of the vertical electric field near the gate caused by the corner, and this fan-out increases for greater V_{g}. As in the symmetric case, we find that the maximum in the velocity distribution is lowered for increasing V_{g}.

Examining the structure of the electric field in Figure 3.57, we find again that the spatial distribution of the field becomes more uniform for increasing V_{g}, with the maximum being found close to the gate, within the oxide. One notes the presence of higher electric field near the drain compared to the symmetric case, as expected. Similar to the hole velocity, the hole density distribution (Figure 3.58) exhibits a fan-out structure due to greater charge accumulation near the source side of the channel. This is, of course, correlated with the lower velocity in these regions. The hole energy (Figure 3.59) reaches its peak value at the end of the gate toward the drain side in the channel. It is at this point that the hole temperature has risen significantly above the lattice temperature. Finally, we observe interesting behavior in the hole current

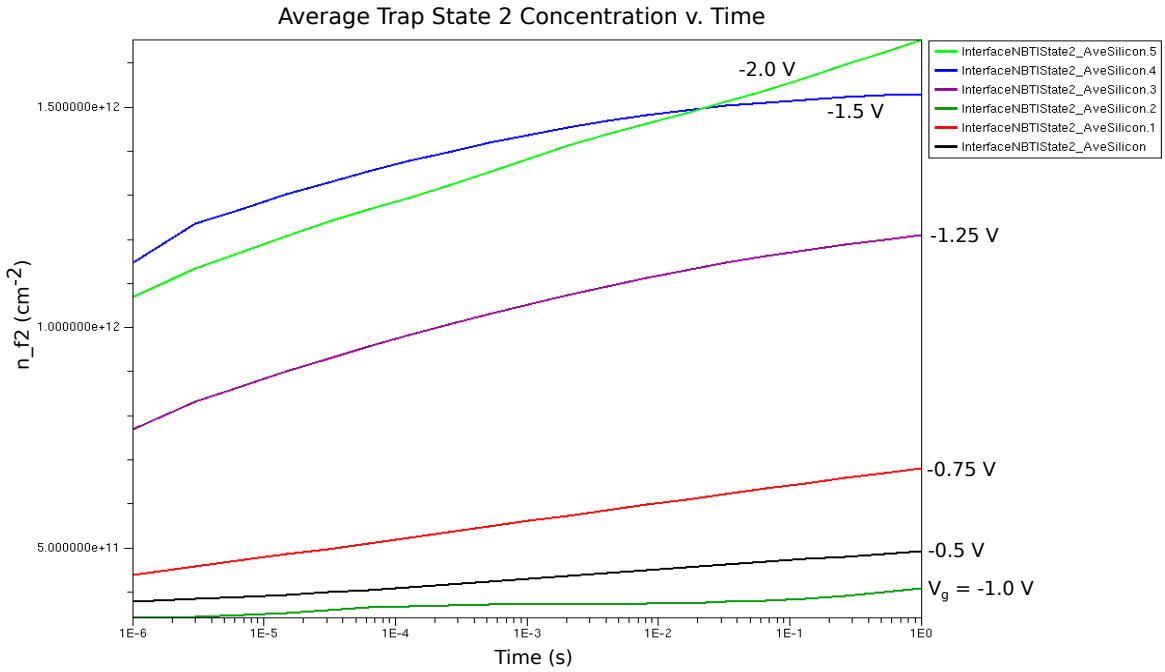


Figure 3.52: Gate voltage variation of $\langle f_2 \rangle$ during asymmetric stress, $T_{\text{ext}} = 350$ K

density for increasing V_{g} (Figure 3.60). We find that the current density becomes strongly peaked near the source-side corner of the gate, while remaining relatively uniform throughout the rest of the channel (excepting the localized minima found for $V_{\text{g}} = -1.0$ V).

Finally, for the asymmetric case, we observe much more pronounced heating of the lattice, and this is attributed to the higher drain bias. The temperature profiles for the standard three gate voltages are shown below (Figure 3.61). Again, we find a hot spot towards the drain side of the channel, except in this case the maximum is on the order of 10 to 30 K above the boundary temperature. That the temperature of the hot spot is not higher is attributed to the Dirichlet thermal boundary conditions. Furthermore, we find that, unlike in the symmetric case, the hot spot widens out for higher V_{g} corresponding to greater hole scattering with the lattice.

We may again raise the question concerning the breaking in the monotonicity of the V_{th} degradation rate observed for symmetric stresses at lower temperatures versus

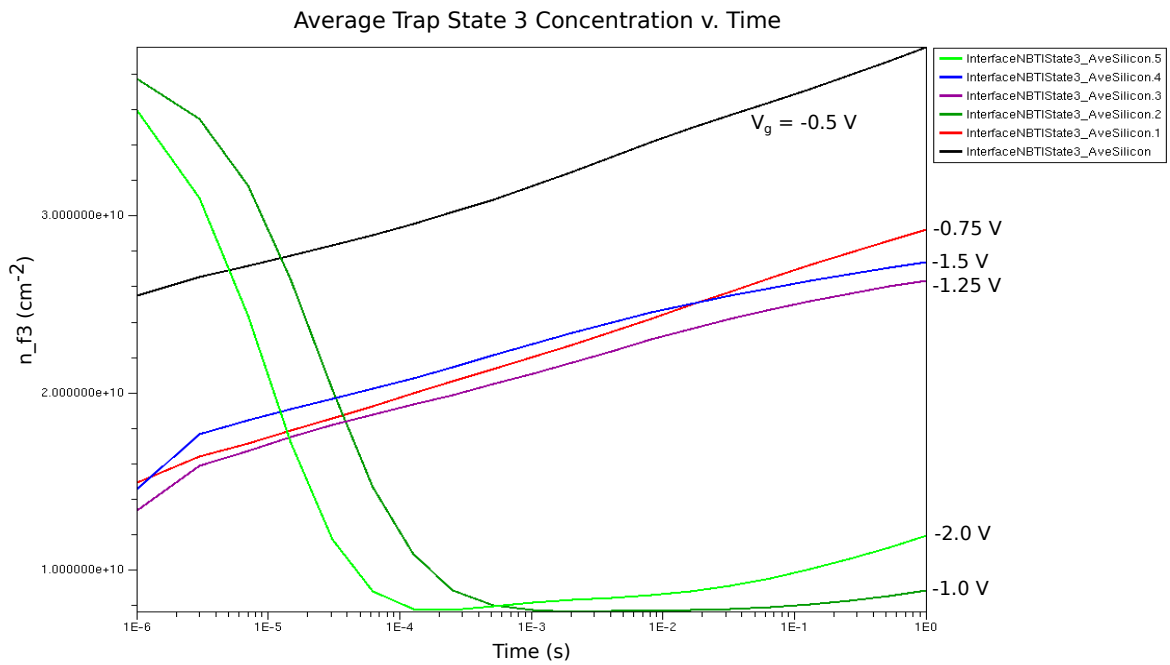


Figure 3.53: Gate voltage variation of $\langle f_3 \rangle$ during asymmetric stress, $T_{\text{ext}} = 350$ K

the monotonic behavior observed for asymmetric stresses and for symmetric stresses at higher temperatures. The peculiar behavior of the trap states under asymmetric stress implies the existence of one or more independent mechanism(s) competing with the NBTI degradation. Indeed, because we observe a monotonic increase in $\Delta V_{\text{th}}(t)$ despite the absence of this behavior for the total trapped charge Q, the degradation in the drain current must be attributed to another mechanism. In particular, the lowering of the maximum hole velocity in the channel and the spatial distribution of the hole current density may indicate the additional source of drain current degradation. The lowering of the maximum hole velocity for increasing V_{g}, or equivalently for increasing electric field, is likely due to velocity overshoot in the channel and the emergence of negative differential resistance [34]. In this case, the lifetime of the charge carrier has exceeded the thermalization time. Indeed, that this velocity overshoot peaks around $V_{\text{g}} = -1.0$ V may explain the higher V_{th} degradation observed at this bias compared to higher gate biases for symmetric stress configurations. Fur-

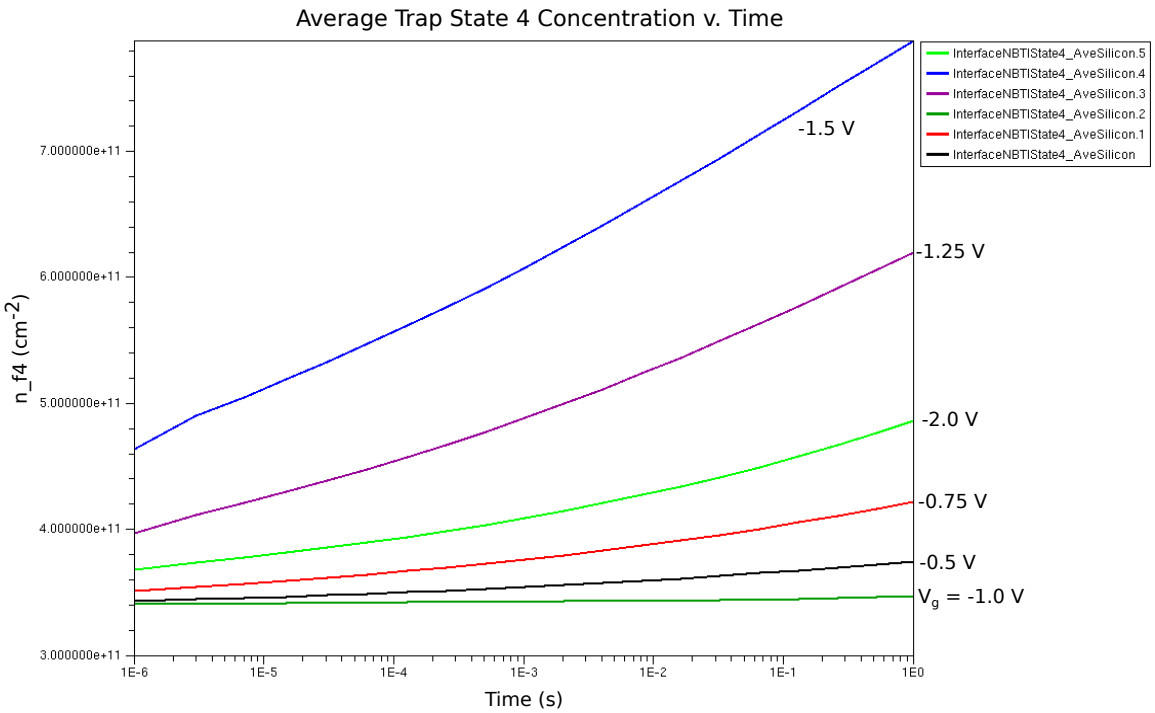


Figure 3.54: Gate voltage variation of $\langle f_4 \rangle$ during asymmetric stress, $T_{\text{ext}} = 350$ K

thermore, it has been well-documented that the hydrodynamic and SHE models tend to overestimate carrier velocities [24, 28, 29]. The apparent velocity overshoot may be eliminated by employing a more accurate subcontinuum simulation based on the BTE [17, 14, 15, 54]. Furthermore, that the maximum hole current density extends across the channel in the symmetric stress configuration, while being highly localized around the source-side edge of the gate in asymmetric stress configurations, may account for the increasing drain current degradation (and hence V_{th} degradation) observed under asymmetric stress but not found under symmetric stress. Indeed, the increase in lattice temperature toward the drain provides further evidence for greater drain current degradation under asymmetric stress. Thus, a combination of velocity overshoot/misestimation and the structures of the hole current density and lattice temperature profiles may contribute to the breaking of the monotonicity in ΔV_{th}.

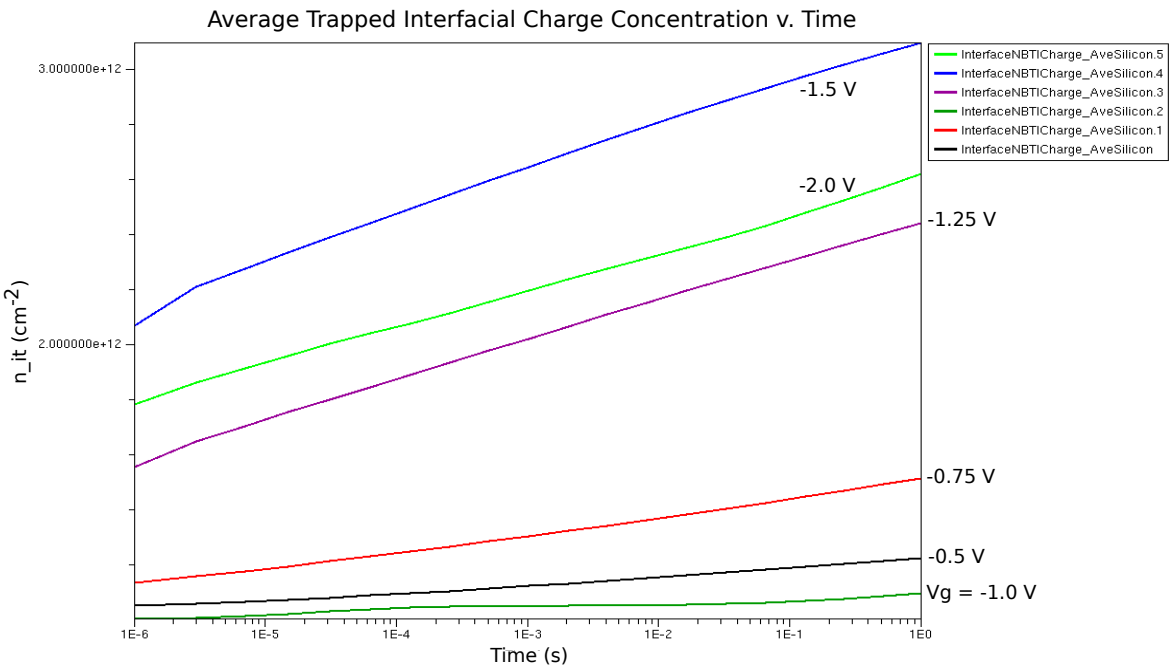


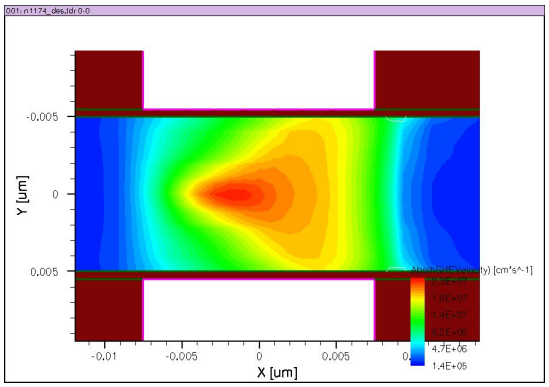
Figure 3.55: Gate voltage variation of Q during asymmetric stress, $T_{\text{ext}} = 350$ K

3.1.3 The Effect of Thermal Boundary Conditions

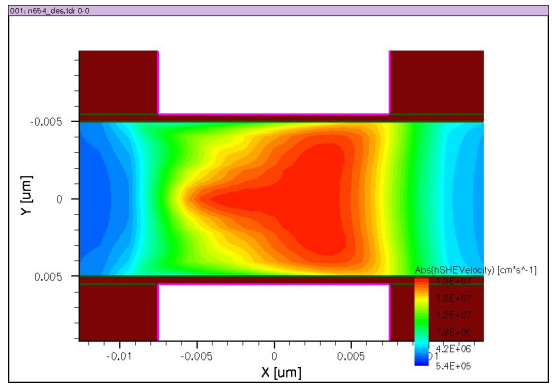
Appropriate boundary conditions are essential for more realistic device modeling, so we consider here the effect of imposing Neumann thermal boundary conditions on the device for a few specific stress cases.[10] (Here, the boundary thermal resistance was set to 0.001 (K·μm)/W). In general, we find very little difference between the distributions for the hole velocity, density, and energy (Figure 3.62(a), Figure 3.63(a), and Figure 3.64(a) respectively)[11] for symmetric stress configurations. This is expected given the very small amount of self-heating attributed to this stress. The situation is different for asymmetric stress configurations. The hole velocity (Figure 3.62(b)) exhibits less fan-out, becoming more localized in the channel beneath the edge of the gate near the drain side. The maximum in hole energy (Figure 3.64(b)) is more spread out and possesses greater concavity. We find similar spreading behavior for

[10]Note that convergence was significantly better for $T_{\text{ext}} = 375$ K than for $T_{\text{ext}} = 350$ K.

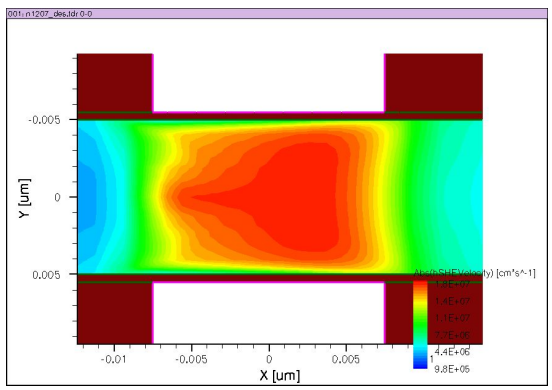
[11]The current density distribution failed to render in TecPlot for this case, but we expect that this too shows little to no change.



(a) $V_{\mathrm{g}} = V_d = -0.5$ V, $T_{\mathrm{ext}} = 350$ K



(b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 350$ K

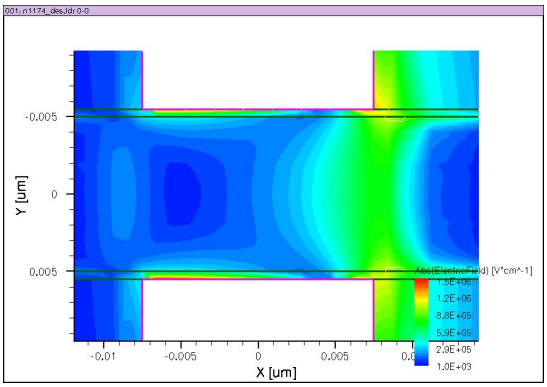


(c) $V_{\mathrm{g}} = V_d = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

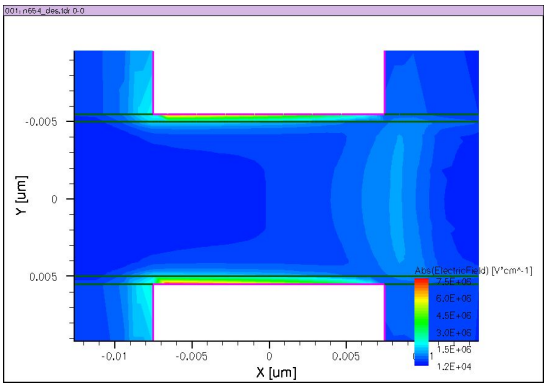
Figure 3.56: Hole velocity $|\boldsymbol{v}_{\mathrm{h}}|$ under asymmetric stress

the minimum region in the hole density (Figure 3.63(b)). For the current density (Figure 3.65), we find that the minimum also becomes more spread out in the center of the channel.

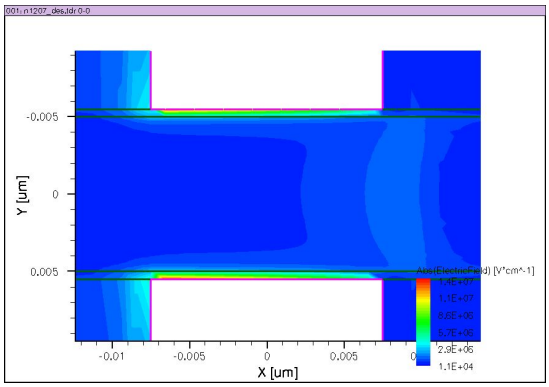
The changes in the characteristic hole distributions observed above follows from the change in the temperature distribution of the lattice under Neumann boundary conditions. Figure 3.66(a) shows the lattice temperature profile under symmetric stress. While the shape of the distribution has changed to accommodate a non-zero thermal resistance at the boundaries, we find that the average temperature of the device rises by about 20 K from the initial 375 K. The difference between maximum and minimum temperatures is again on the order of 0.1 K.



(a) $V_{\mathrm{g}} = V_d = -0.5$ V, $T_{\mathrm{ext}} = 350$ K



(b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 350$ K

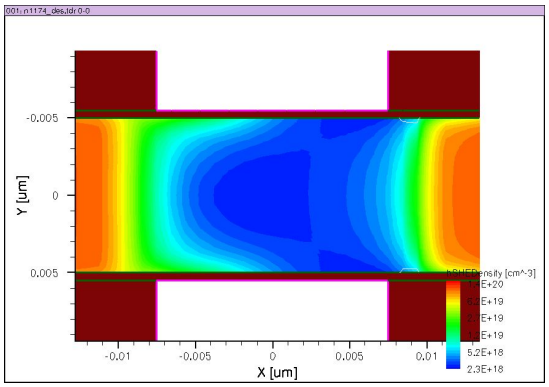


(c) $V_{\mathrm{g}} = V_d = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

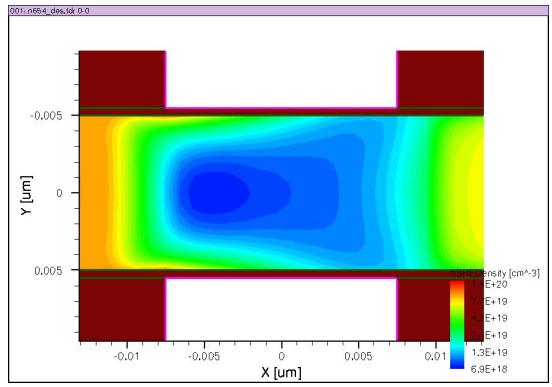
Figure 3.57: Electric field $|\boldsymbol{E}|$ under asymmetric stress

Figure 3.66(b) shows the lattice temperature profile for an asymmetric stress configuration. In this case we find a more localized hot spot corresponding, again, to the greater degree of charge scattering with the lattice toward the drain side of the channel. However, we find a significant rise in the average temperature of the device on the order of 750 K greater than the initial 375 K. Furthermore, we find that the hot spot possesses a temperature roughly 100 K higher than the minimum temperature of the device. Neumann thermal boundary conditions thus allow for significant increases in the FinFET lattice temperature since the temperature is not fixed at the boundaries as it is for Dirichlet boundary conditions.

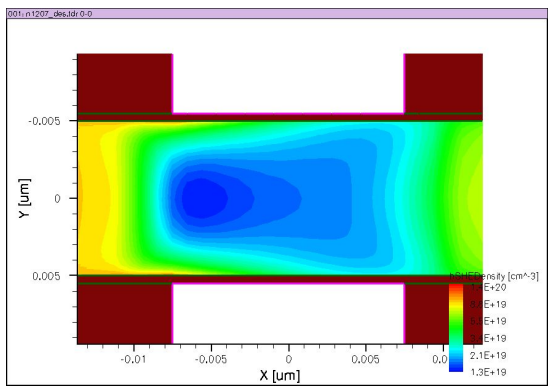
For completeness, we give a brief comparison between the SHE model and the hydrodynamic model in simulating the 2D FinFET. Figure 3.67 shows the difference in



(a) $V_{\mathrm{g}} = V_d = -0.5$ V, $T_{\mathrm{ext}} = 350$ K



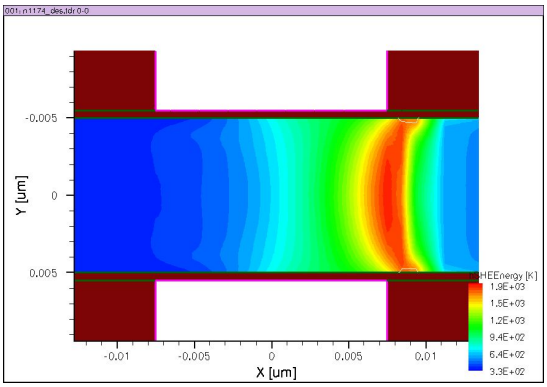
(b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 350$ K



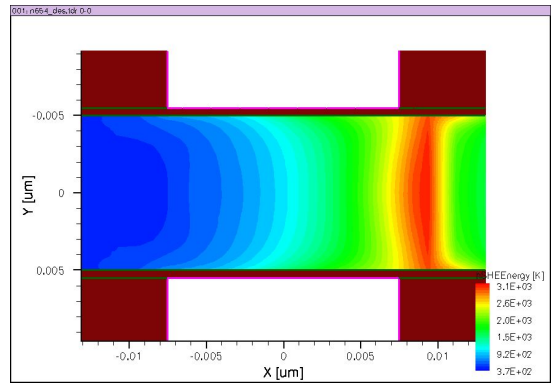
(c) $V_{\mathrm{g}} = V_d = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

Figure 3.58: Hole density n_{h} under asymmetric stress

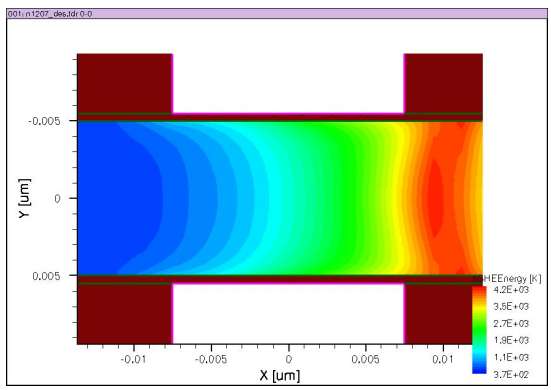
ΔV_{th} following a symmetric NBTI stress for the SHE and hydrodynamic models. In Figure 3.68(a), we find that the hole velocity is qualitatively similar to the SHE case for symmetric stress, while Figure 3.68(b) does not account for the peak in velocity near the drain under asymmetric stress. Figure 3.69 shows similar distributions for the hole density computed by the SHE and hydrodynamic models for both the symmetric and asymmetric stress. For symmetric stress, we find that the hydrodynamic model (Figure 3.70(a)) gives a very different result for the hole temperature (which is equivalent to the hole energy computed by the SHE) than the SHE model, and this points to the fundamental difference in the mathematical structure between the two models. The SHE and hydrodynamic models show better agreement on hole temperature for asymmetric stress (Figure 3.70(a)). Figure 3.71 shows the hole current



(a) $V_{\mathrm{g}} = V_d = -0.5$ V, $T_{\mathrm{ext}} = 350$ K



(b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 350$ K



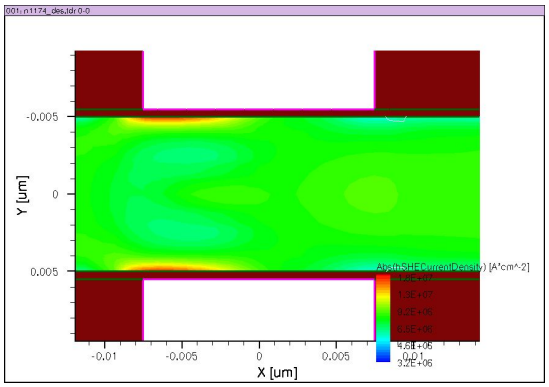
(c) $V_{\mathrm{g}} = V_d = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

Figure 3.59: Hole energy ϵ_{h} under asymmetric stress

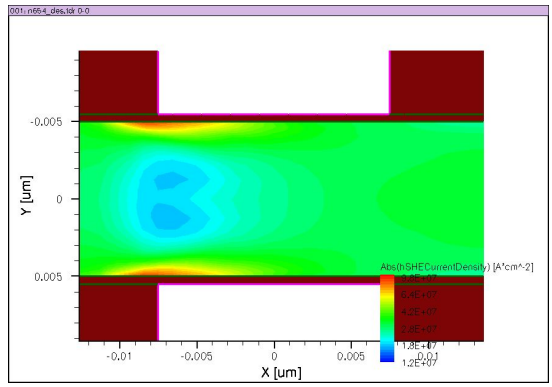
density for the hydrodynamic model with symmetric and asymmetric stress, and we observe an exaggeration in the region of local minima compared to the SHE model in both cases. Finally, we find similar lattice temperature profiles for symmetric and asymmetric stresses (Figure 3.72), except that the hot spot in the hydrodynamic model has spread further into the oxynitide spacers.

3.1.4 Simulation Convergence and Numerical Error

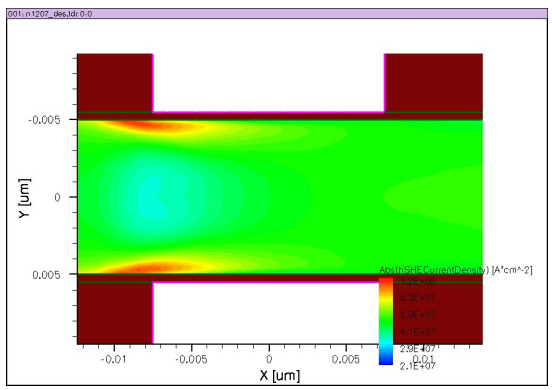
The numerical routine implemented by Sentaurus TCAD solves a coupled set of discretized device equations self-consistently using the backward Euler (BE) method as an implicit discretization scheme [77]. Thus, the equations are solved at a given time t_i with a given time step h_i, and the simulation proceeds to the next point in



(a) $V_{\mathrm{g}} = V_d = -0.5$ V, $T_{\mathrm{ext}} = 350$ K



(b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 350$ K

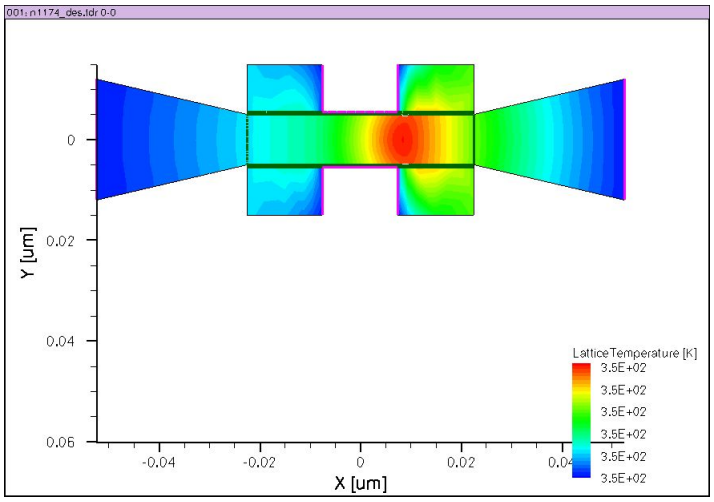


(c) $V_{\mathrm{g}} = V_d = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

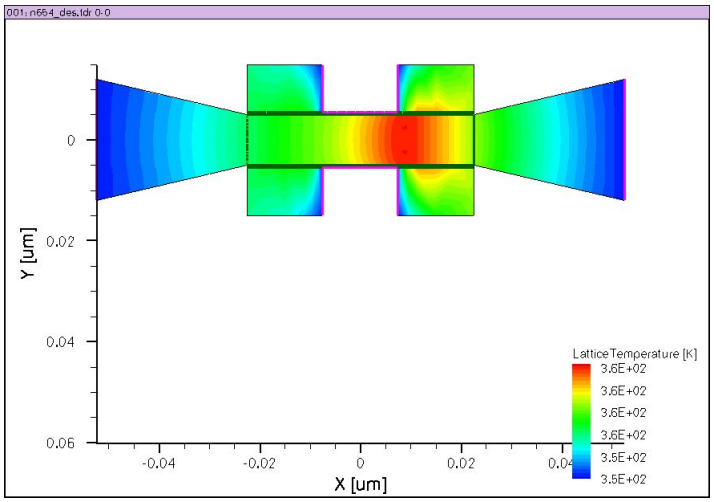
Figure 3.60: Hole current density $|\boldsymbol{J}_{\mathrm{h}}|$ under asymmetric stress

time $t_{i+1} = t_i + h_i$ and calculates a new time step $h_{i+1} > h_i$ if the solution converges. If the simulation does not converge at a given time, the equations are solved for a smaller time step $h_i := h_i/2$. If the simulation fails to converge for a minimum time step, the loop will break, indicating that the run has failed. Thus, for a successfully converged solution over the time domain, we find that the time step will increase as the simulation loops through the solver.

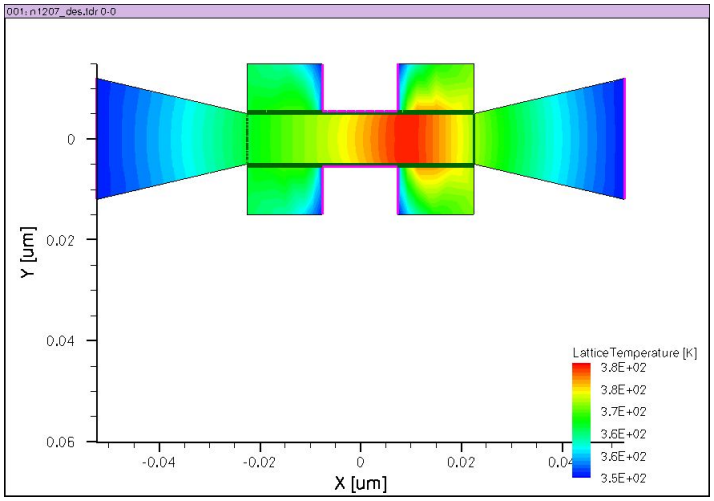
While the time-stepping of the quasistationary and transient simulations is regulated by a maximum number of iterations and a minimum time step $h_{\min}$, the convergence of the solution at each point in time t_i is controlled by both absolute and relative error criterion for each calculated quantity. A solution for the quantity f_{i+1}^n is converged at node n for an updated time t_{i+1} for



(a) $V_{\mathrm{g}} = V_d = -0.5$ V, $T_{\mathrm{ext}} = 350$ K

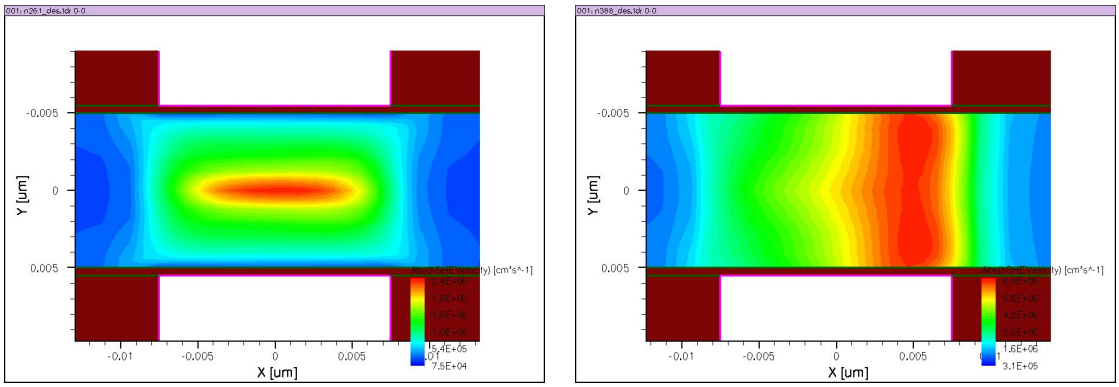


(b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 350$ K



(c) $V_{\mathrm{g}} = V_d = -1.5$ V, $T_{\mathrm{ext}} = 350$ K

Figure 3.61: Lattice temeperature T_{L} under asymmetric stress



(a) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 375$ K (b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K

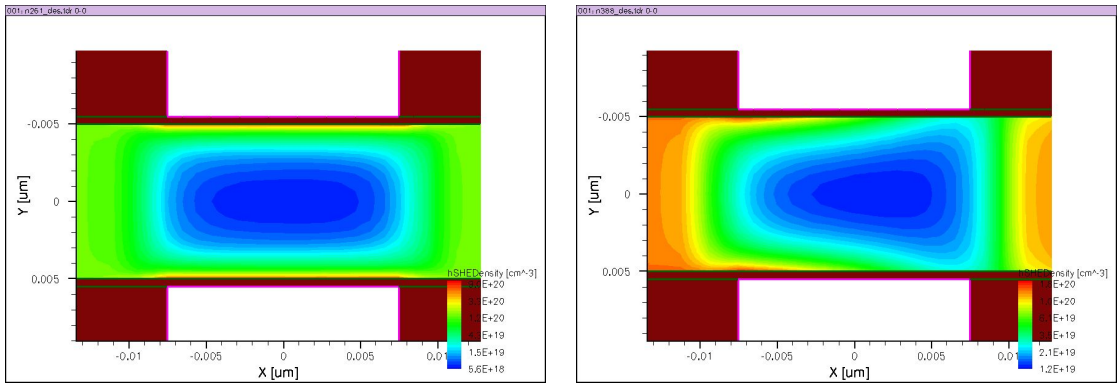
Figure 3.62: Hole velocity $|\boldsymbol{v}_{\mathrm{h}}|$: symmetric versus asymmetric stress with Neumann boundary conditions

$$\frac{\left|(f_{i+1}^n - f_i^n)/\lambda\right|}{\epsilon_{\mathrm{R}}\left|f_{i+1}^n/\lambda\right| + \epsilon_{\mathrm{A}}} < 1 \ , \tag{3.1}$$

where f_i^n is the quantity value at the previous time, λ is a scaling factor, ϵ_{R} is the relative error, and ϵ_{A} is the absolute error for the specified quantity. The scaling factor λ is determined by a reference value $f_{\mathrm{ref}} = \lambda\epsilon_{\mathrm{A}}/\epsilon_{\mathrm{R}}$ for the calculation. The reference value f_{ref}, or equivalently the scaling factor λ, is calibrated against simulation and experimental data for silicon, and it serves to provide stability to the convergence of the BE method. In this work, we used the default reference values and absolute errors ($\epsilon_{\mathrm{A}} \leq 10^{-3}$) calibrated for Sentaurus TCAD for each calculated quantity. The relative error ϵ_{R} is determined by the number of digits D, and thus the precision, used in the solution: $\epsilon_{\mathrm{R}} = 10^{-D}$. In this work, $D = 6$. Convergence of the simulation for the set of device equations across the entire device is then given by an average over the error measure at each node:

$$\frac{1}{\epsilon_R N}\sum_{n,\xi}\frac{\left|f_{i+1}^n(\xi) - f_i^n(\xi)\right|}{\left|f_{i+1}^n(\xi)\right| + f_{\mathrm{ref}}} < 1 \tag{3.2}$$

Here we have employed the notation $f_i^n(\xi)$ to denote the solution to equation ξ at node n and iteration point i, and N is the total number of device equations multiplied



(a) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 375$ K (b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K

Figure 3.63: Hole density n_{h}: symmetric versus asymmetric stress with Neumann boundary conditions

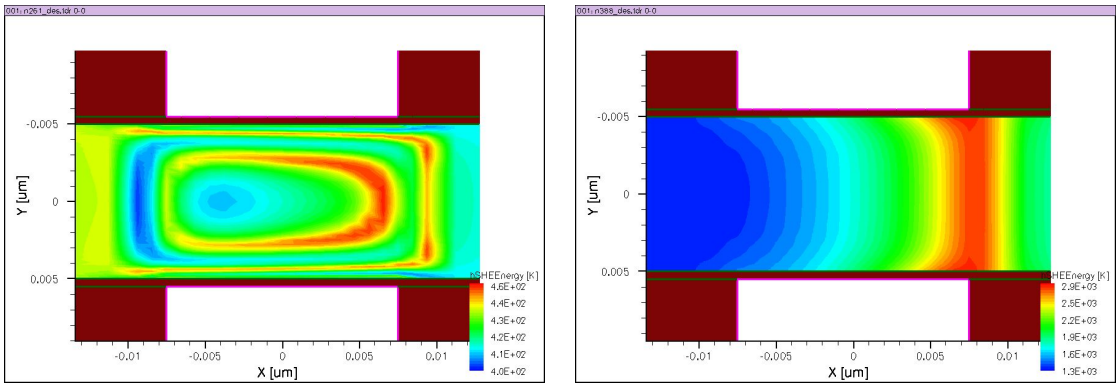
by the total number of nodes for the finite element mesh of the device. This error criterion requires the mutual convergence of each equation solution with respect to the full set of device equations. For a converged solution using the BE method, the next time step is calculated from the L^2 norm of the error:

$$h_{i+1} = \frac{h_i}{\sqrt{r}} \ , \tag{3.3}$$

where

$$r = \sqrt{\frac{1}{\epsilon_{\mathrm{R}} N} \sum_{n,\xi} \left(\frac{\left| f_{i+1}^n(\xi) - f_i^n(\xi) \right|}{\left| f_{i+1}^n(\xi) \right| + f_{\mathrm{ref}}} \right)^2} \ . \tag{3.4}$$

In Sentaurus TCAD, convergence data is only stored temporarily as required to calculate the relative error in the Newtonian iteration, in order to reduce the memory requirements of the simulation. However, we can measure the convergence quality of the simulation by examining the quasistationary ramps of device parameters, and this is typically done in practice for TCAD simulation. Figure 3.73(a) and Figure 3.73(b) show the ramp of gate and drain biases during the quasistationary part of an asymmetric stress simulation on the 2D FinFET. Here, the gate and drain biases are ramped to -2.0 V at a low ambient temperature (300 K). This represents an extreme



(a) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 375$ K (b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K

Figure 3.64: Hole energy ϵ_{h}: symmetric verses asymmetric stress with Neumann boundary conditions

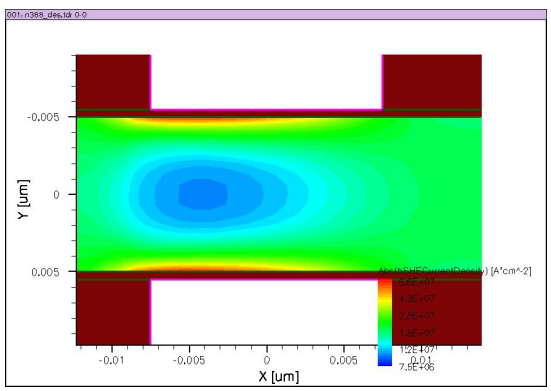
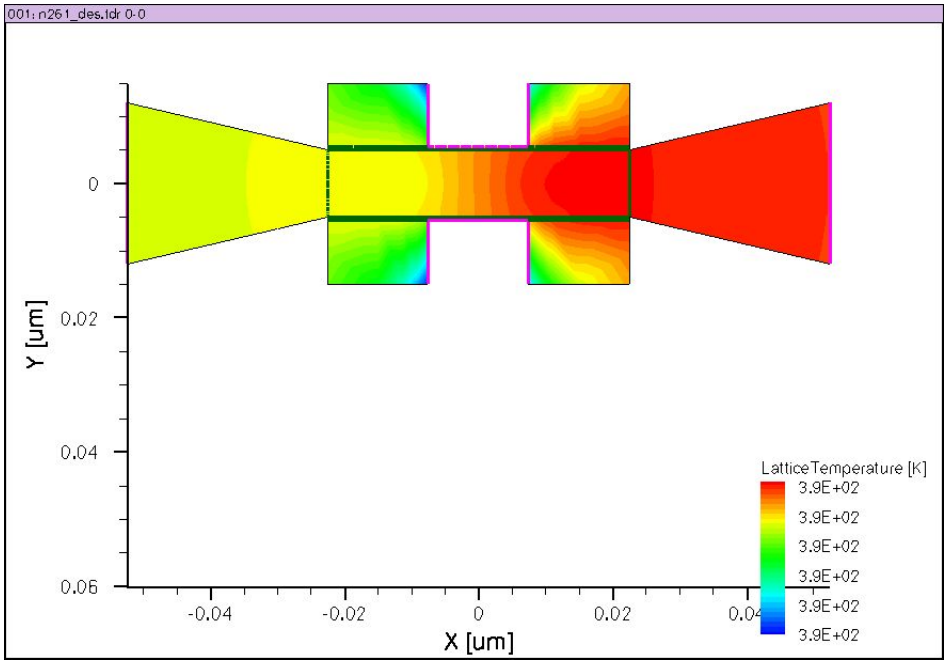


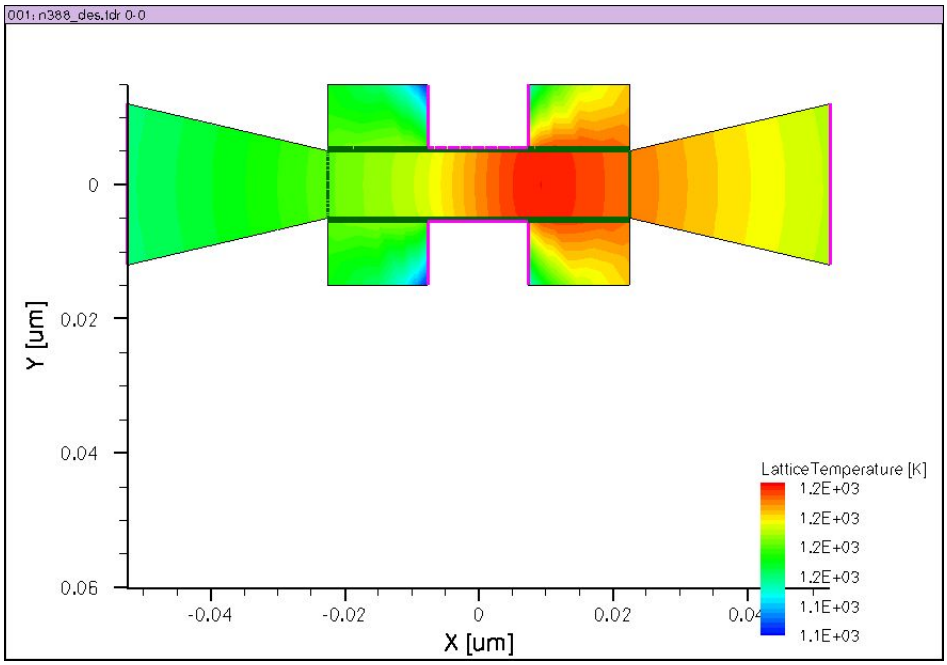
Figure 3.65: Hole current density n_{h} under asymmetric stress ($V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K) with Neumann boundary conditions

case where mesh refinement was required to improve convergence. Figure 3.73(a) shows an instance where the simulation has failed to converge, iterating over a virtual time parameter $0 < t_i < 1$. The time separation between successive data points decreases as the loop fails to find a converged solution. This is the typical case for failed convergence in TCAD. Figure 3.73(b) shows an instance where the solution has converged at each point in time, successfully ramping the gate and drain to -2.0 V. The final result is converged to six digits as required by the relative error.

The BE discretization method approximates the time-evolution of the system up to first order. For a general evolution equation,



(a) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 375$ K



(b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K

Figure 3.66: Lattice temperature T_{L}: symmetric versus asymmetric stress with Neumann conditions

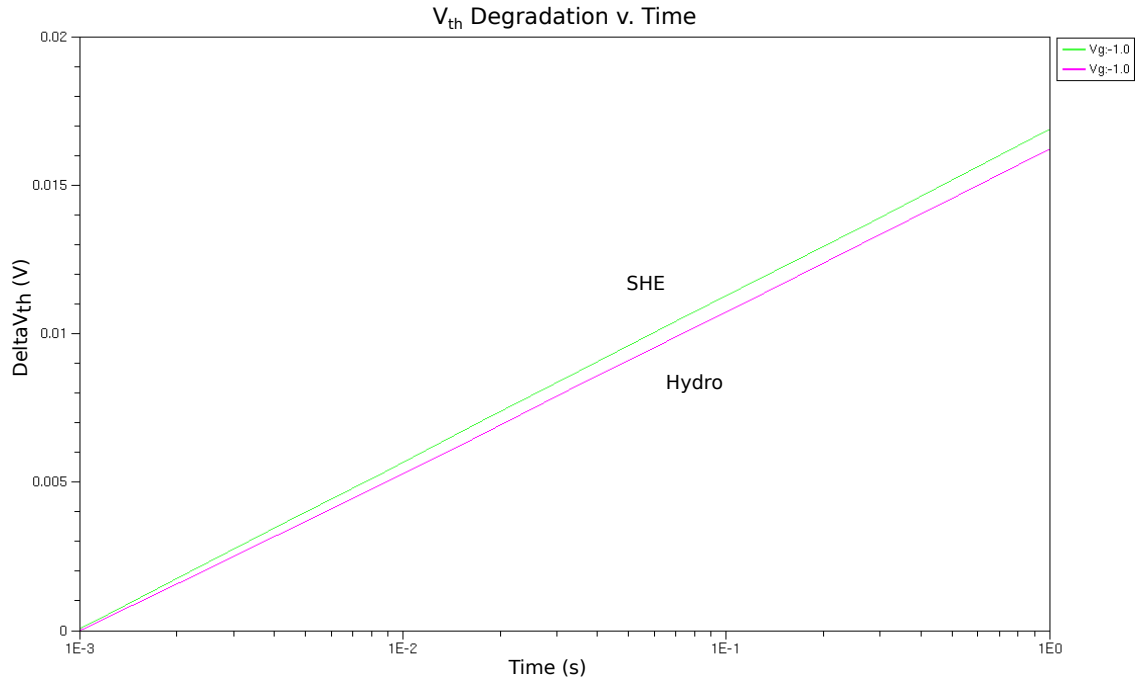


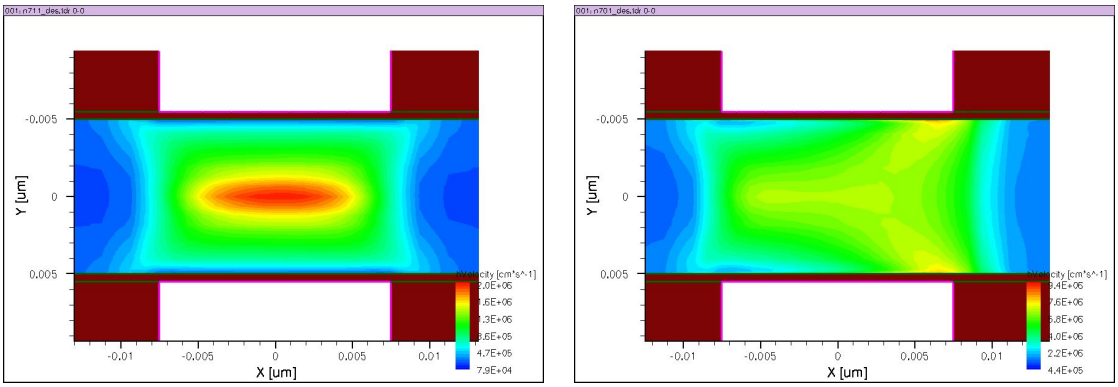
Figure 3.67: Relative $\Delta V_{\text{th}}(t)$ for SHE and Hydrodynamic simulations ($V_{\text{g}} = -1.0$ V, $T_{\text{ext}} = 325$ K)

$$\frac{d}{dt} f[q(t)] + g[q(t), t] = 0 \ , \tag{3.5}$$

this discretization is written as

$$f(t_i + h_i) + h_i g(t_i + h_i) = f(t_i) \ . \tag{3.6}$$

Thus, there is an intrinsic error associated with the truncation of the approximation and the subsequent time-stepping based on this approximation. While the error criterion (3.2) regulates the numerical convergence of the Newton iterations during the simulation, it does not account for this error in the time-stepping. Rather, the time step is only controlled by the convergence of the solutions to the device equations. This local truncation error is improved by using a more accurate discretization scheme, such as the trapezoidal rule/backward differentiation formula (TRBDF), although at the cost of simulation run-time.

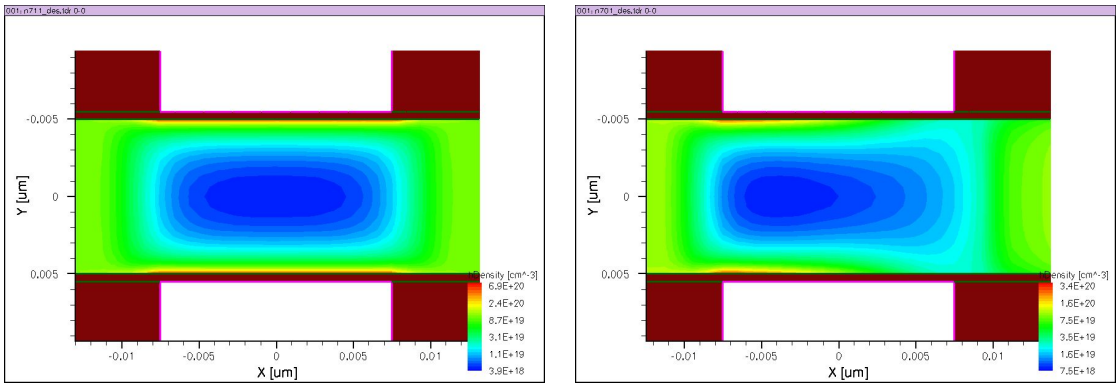


(a) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 375$ K (b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K

Figure 3.68: Hole velocity $|\boldsymbol{v}_{\mathrm{h}}|$: symmetric versus asymmetric stress with Hydrodynamic model

The accuracy of the solutions to the device equations computed in TCAD depends strongly on the finite element mesh used to emulate the device structure. Indeed, highly localized or rapidly spatially-fluctuating phenomena pose a significant challenge to proper finite element modeling. Sentaurus TCAD uses a Delaunay axis-aligned meshing algorithm to generate a mesh refined for device region boundaries and material interfaces [78]. Further refinement of the mesh improves the numerical convergence of the simulation as it reduces the interpolation error between adjacent nodes. In this work, the mesh was refined by including a higher number of nodes along material boundaries and corners in order to improve the convergence of the simulation. While this extends the total run-time of the simulation, fewer iterations at each time step are required for a converged solution. In this work, we found that half the iterations at each time step were required after mesh refinement than before.

The two-stage NBTI model possesses inherent statistical error, as the average change in positive charge at the Si/SiO_2 interface is determined using a random sampling technique for each interface node. Indeed, the simulation creates N_s random configurations for each interface node from the set of random variables (2.98), and the average over each calculated quantity f_i at each node is given generally by



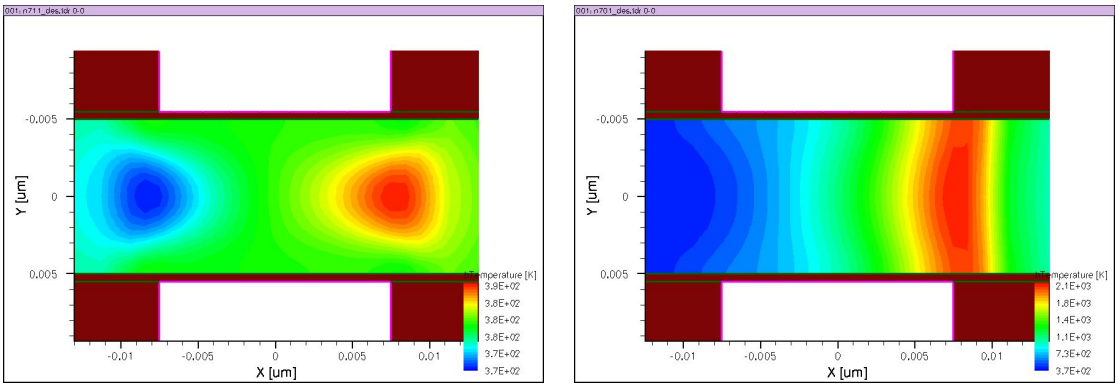
(a) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 375$ K (b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K

Figure 3.69: Hole density n_{h}: symmetric versus asymmetric stress with Hydrodynamic model

$$\langle f \rangle = \frac{1}{N_s} \sum_{i=1}^{N_s} f_i \tag{3.7}$$

In this work, $N_s = 1000$, and the $\{f_i\}$ are chosen according to the distribution of each variable in (2.98). Often, we find that spatial averaging of the NBTI trap states across the interfaces of the device require several reductions in the time step at each iteration point in order to satisfy the convergence criterion (3.2). This is indicative of statistical noise in the two-stage NBTI model, which dominates the initial transient behavior of the trap states for $t < 10^{-3}$ s (e.g., Figure 3.24). We found that increasing the initial time step from 10^{-15} to 10^{-6} s effectively halved the number of iterations required for convergence at each point in time. However, since the degradation in threshold voltage is only computed for $t > 10^{-3}$ s while the device is stressed, the results for $\Delta V_{\mathrm{th}}(t)$ are unaffected for smaller initial time steps as long as the solution has converged.

Convergence is consistently better for symmetric stress configurations than for asymmetric stress configurations. After refining the mesh, we found that between symmetric and asymmetric configurations with the same external parameters (e.g., T_{ext}, V_{g}, initial trap precursor density), the calculated lattice and hole temperatures

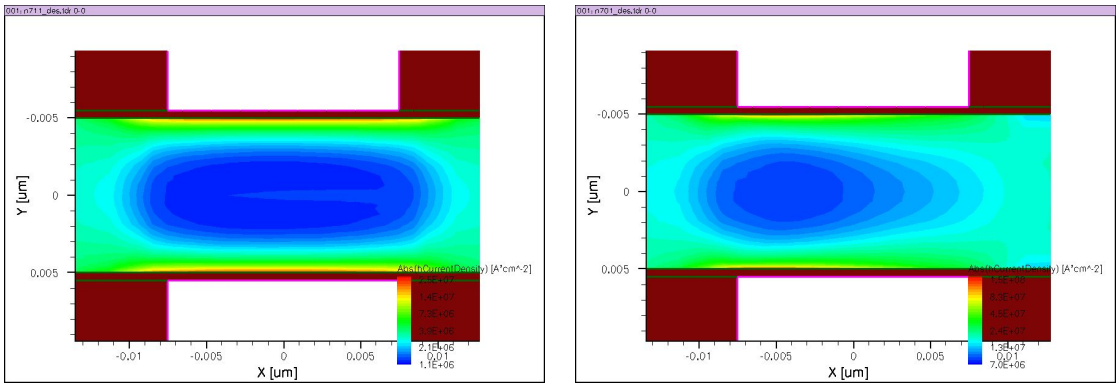


(a) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 375$ K (b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K

Figure 3.70: Hole temperature T_{h}: symmetric versus asymmetric stress with Hydrodynamic model

and hole current exhibited higher fluctuations for asymmetric configurations, and the simulation required twice the number of iterations at each time step before converging according to (3.2). Indeed, this high degree of interaction between the lattice and hole temperatures and the ramping of the gate and drain biases often requires several iterations at each time point in order to satisfy the electrical and thermal boundary conditions. Furthermore, we find that the absolute limit for convergence of the asymmetric stress configuration is for $V_{\mathrm{g}} < -2.0$ V and $T_{\mathrm{ext}} < 325$ K. For symmetric configurations, on the other hand, this limit is relaxed to $V_{\mathrm{g}} < -2.25$ V and $T_{\mathrm{ext}} < 275$ K. Nevertheless, our results are converged for external parameters within the appropriately defined stress range and for both the symmetric and asymmetric stress configurations by the criterion (3.2). That this is the case is demonstrated by the fact that our transient simulation computes solutions to NBTI trap states for the entire required run-time ($t_{\mathrm{f}} = 1$ s), since the Newtonian iterations would break and end the simulation otherwise.

We find that numerical convergence is particularly sensitive to thermal boundary conditions. We observed that the simulation converges faster and for a broader range of stress conditions for Dirichlet thermal boundary conditions compared to Neumann

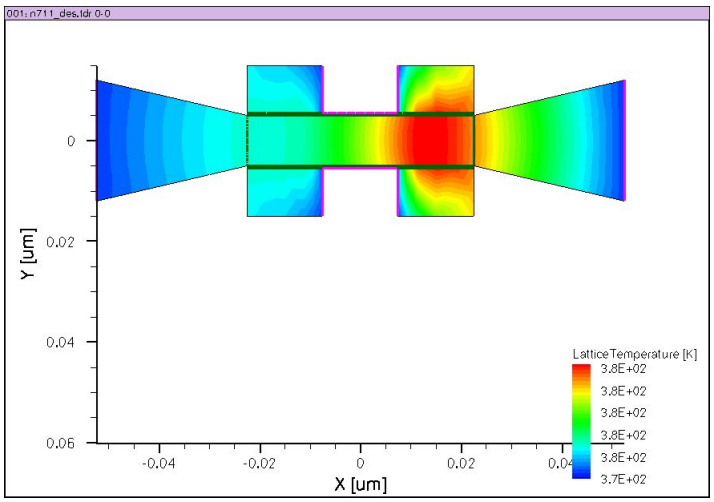


(a) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K (b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K

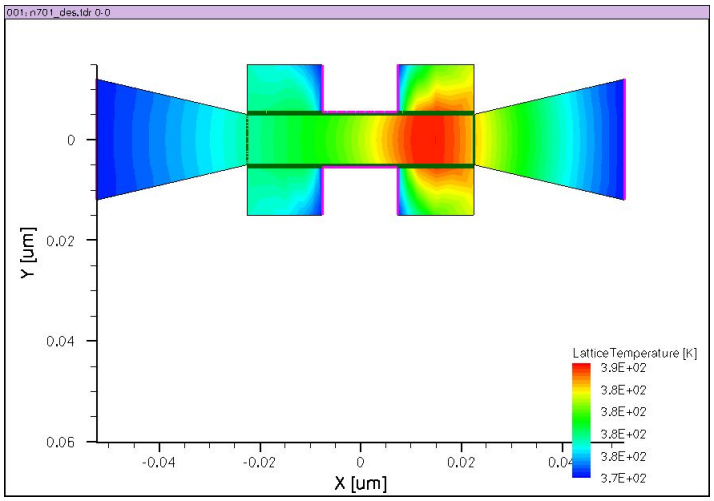
Figure 3.71: Hole current density $|\boldsymbol{J}_{\mathrm{h}}|$: symmetric versus asymmetric stress with Hydrodynamic model

thermal boundary conditions. This can be attributed to the fact that Neumann boundary conditions do not fix the temperature at the boundaries of the device, requiring the simulation to solve for the temperature along the boundary and to propagate the solution at each boundary node into the bulk of the device at each iteration. In particular, we examined four test cases in order to characterize the effect of the thermal boundary conditions on the convergence of the simulation after mesh refinement. For a symmetric configuration with fixed gate voltage $V_{\mathrm{g}} = -1.5$ V and ambient temperature $T_{\mathrm{ext}} = 350$, we ran the simulation first with Dirichlet boundary conditions and then with Neumann boundary conditions with thermal resistances of $R_{\mathrm{th}} = 0.1$, 0.01, and 0.001 K$\cdot\mu$m)/W. The simulation with Neumann boundary conditions failed to converge for $R_{\mathrm{th}} = 0.1$ K$\cdot\mu$m)/W, while the number of iterations required for convergence at each time step decreased when R_{th} was decreased from 0.01 to 0.001 K$\cdot\mu$m)/W. Dirichlet boundary conditions showed a further improvement for the number of iterations at each time step, requiring on average 33 % of the iterations for Neumann boundary conditions with $R_{\mathrm{th}} = 0.001$ K$\cdot\mu$m)/W.

Furthermore, we found that convergence under a broader range of stress configurations is improved for higher boundary temperatures for both Dirichlet and Neumann

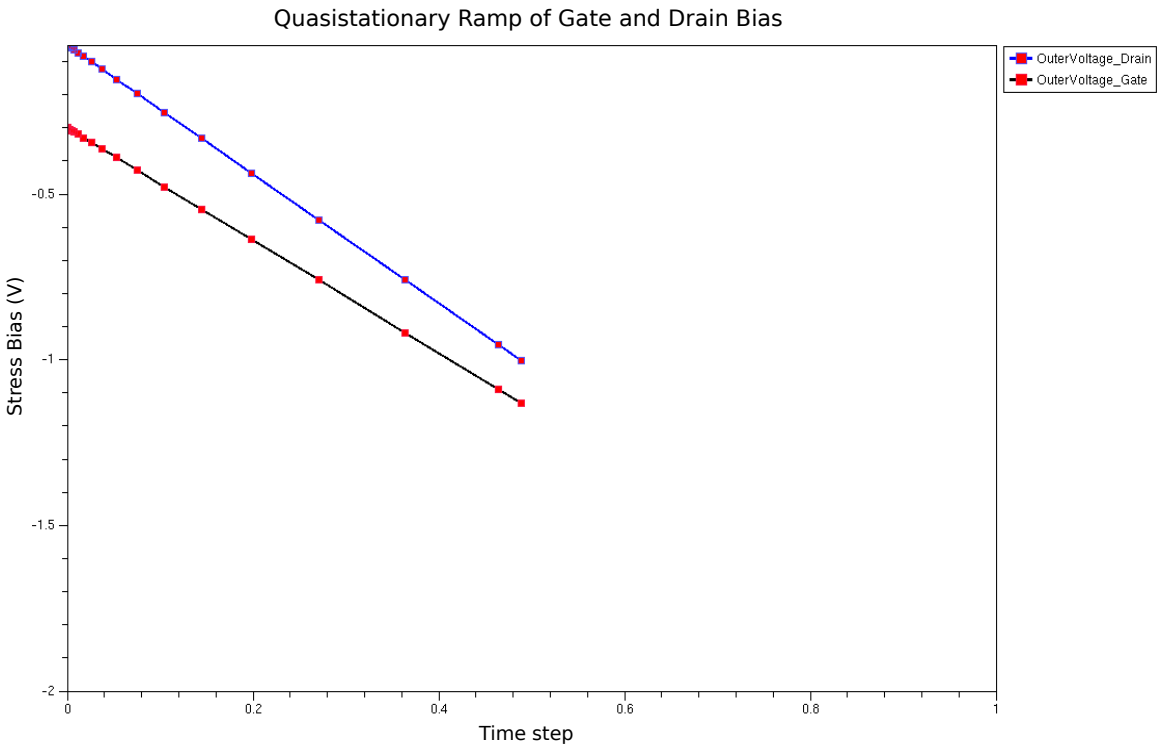


(a) $V_{\mathrm{g}} = -1.0$ V, $T_{\mathrm{ext}} = 375$ K

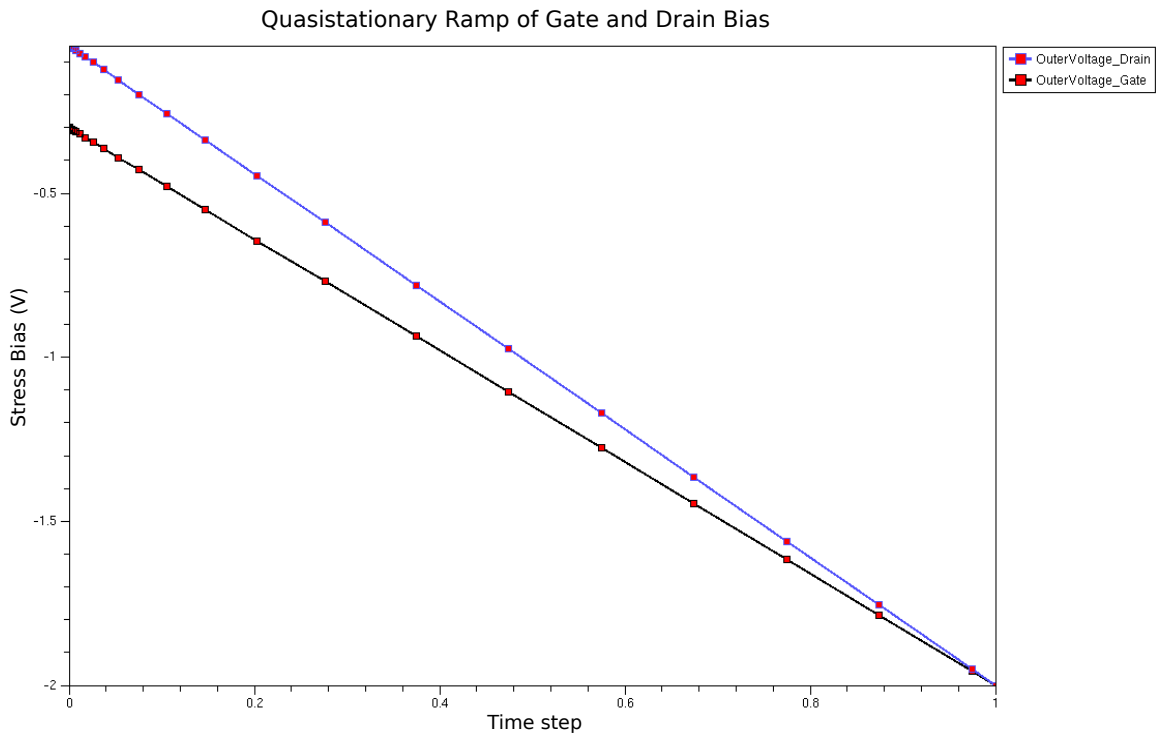


(b) $V_{\mathrm{g}} = V_d = -1.0$ V, $T_{\mathrm{ext}} = 375$ K

Figure 3.72: Lattice temperature T_{L}: symmetric versus asymmetric stress with Hydrodynamic model



(a) Failed convergence during quasistationary ramp



(b) Successful convergence of quasistationary ramp

Figure 3.73: Failed versus successful convergence during quasistationary ramp for 2D FinFET NBTI stress simulations ($V_g = -2.0$ V)

boundary conditions. Why this is the case remains an open question. For a symmetric stress configuration with Dirichlet boundary conditions, which represents the best case scenario, we find that for low temperatures a high degree of fluctuation in the solutions causes the simulation to reduce the time step at the same time point until the simulation breaks. This often occurs early (i.e., for $t_i \approx 10^{-5}$ s $\ll t_f$) in the simulation. Thus, we have restricted ourselves to examining stress configurations for $T_{\text{ext}} \geq 300$ K. As T_{ext} is increased, the number of iterations required for convergence at each time point decreases. For $T_{\text{ext}} > 500$ K, the simulation only iterates once at each point in time.

In general, the stress and relaxation NBTI simulations for each simulation case (for 2D and 3D FinFETs and the 2D MOSFET) converge better for lower biases and higher temperatures. More consistent convergence for a variety of boundary conditions and model parameters can be achieved by refining and tightening the finite element mesh (which is done by adding more node points along material boundaries and corners and rerunning the meshing algorithm), although this often significantly increases the simulation time due to the additional number of nodes. As we have noted above, due to noise in the two-stage NBTI model, convergence is further improved by taking a larger initial time step in the transient portion of the stress and relaxation simulations. The symmetric stress configuration converges much faster than the asymmetric stress configuration. Furthermore, the choice of thermal boundary conditions greatly affects the convergence and run time of both conditions. The simulation will converge faster and for a broader range of stress conditions for Dirichlet thermal boundary conditions than for Neumann thermal boundary conditions. Again, for the results presented in this work, each calculated quantity is converged to at least six digits as required by 3.2.

3.2 Case 2: 2D pMOSFET

As a consistency check for the 2D FinFET case, we have run the NBTI simulation on a 2D SOI pMOSFET structure with gate length $L_g = 40$ nm for symmetric stress configurations at lower temperatures. The 2D structure with finite element mesh and doping profile are shown in Figure 3.74. Again, electrical/thermal boundaries are indicated by the pink edges, and the gate, source and drain regions have been replaced by the contact edges. Furthermore, the orientation of source and drain are reversed compared to the 2D FinFET structure, with the drain on the left and source on the right. The results for threshold voltage degradation are given in Figure 3.75, Figure 3.76, and Figure 3.77. Note that convergence failed for $T_{\text{ext}} = 325$ K. Furthermore, the solution for this 2D MOSFET case employs the SHE model in the device equations.

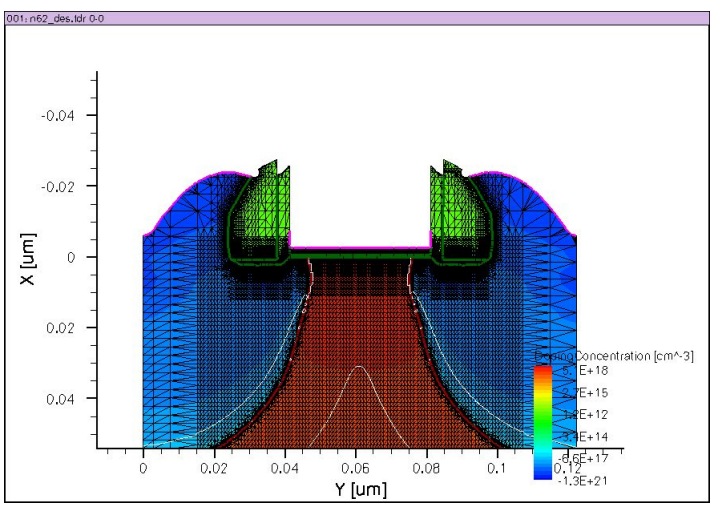


Figure 3.74: Finite element mesh and doping concentration for 2D MOSFET structure

Again, we observe a break in the expected monotonicity of the rate of threshold voltage degradation with increasing gate voltage. Furthermore, we find that the separation between successive ΔV_{th} curves has increased compared to the 2D FinFET case. An examination of the trap states 1-4 and the total trapped charge at $T_{\text{ext}} = 375$ K (Figure 3.78, Figure 3.79, Figure 3.80, Figure 3.81, and Figure 3.82) reveals regular

behavior. Indeed, f_2, f_4 and Q increase with increasing V_g, while f_1 and f_3 respectively decrease.

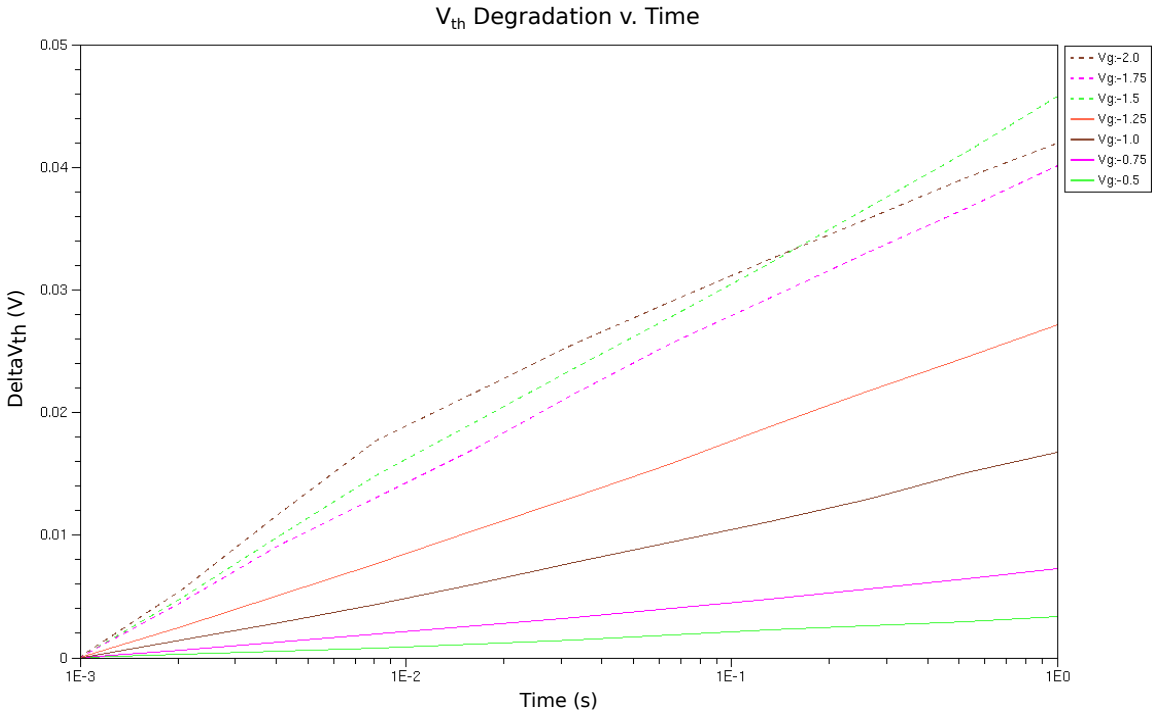


Figure 3.75: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 300$ K

It is worth noting that the hole velocity develops a maximum in the channel far below the gate for high V_g (Figure 3.83(a)). The electric field exhibits similar corner effects near the gate (Figure 3.83(b)) compared to the 2D FinFET. Furthermore, we find a similar resonant-like distribution for the hole energy (Figure 3.83(d)) as found in the 2D FinFET. Finally, we find a small hot spot in the lattice temperature near the drain on the order 10 K higher than the boundary temperature.

3.3 Case 3: 3D pFinFET

Having built up our model for the 2D FinFET, we extend our analysis to a 3D FinFET structure. This structure was supplied by IBM's TCAD team and adapted for compatibility with Sentaurus TCAD. We are, thus, restricted in the material and device parameters we can supply for this structure, although it is generally analogous

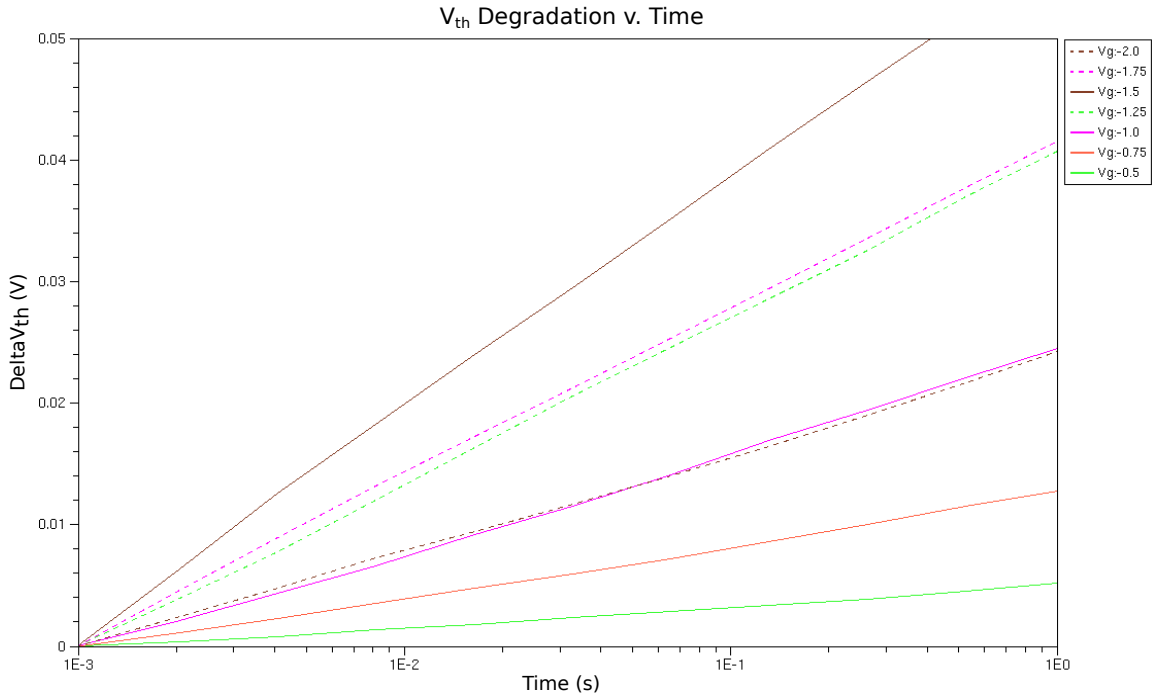


Figure 3.76: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 350$ K

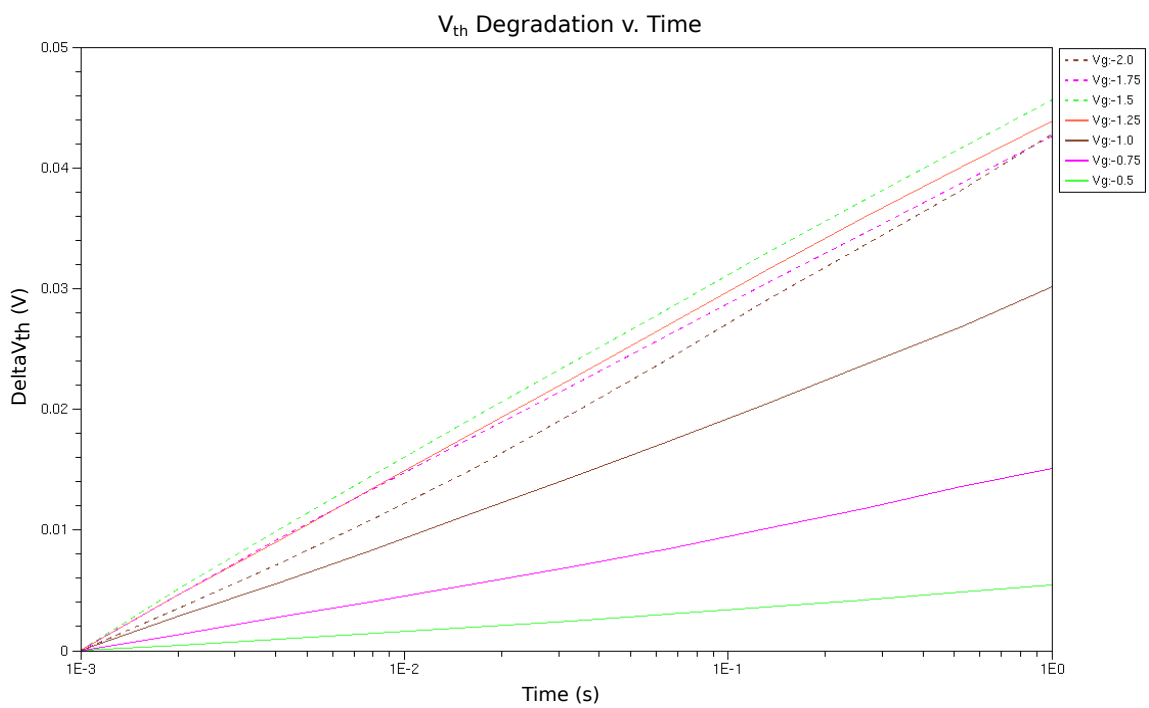


Figure 3.77: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 375$ K

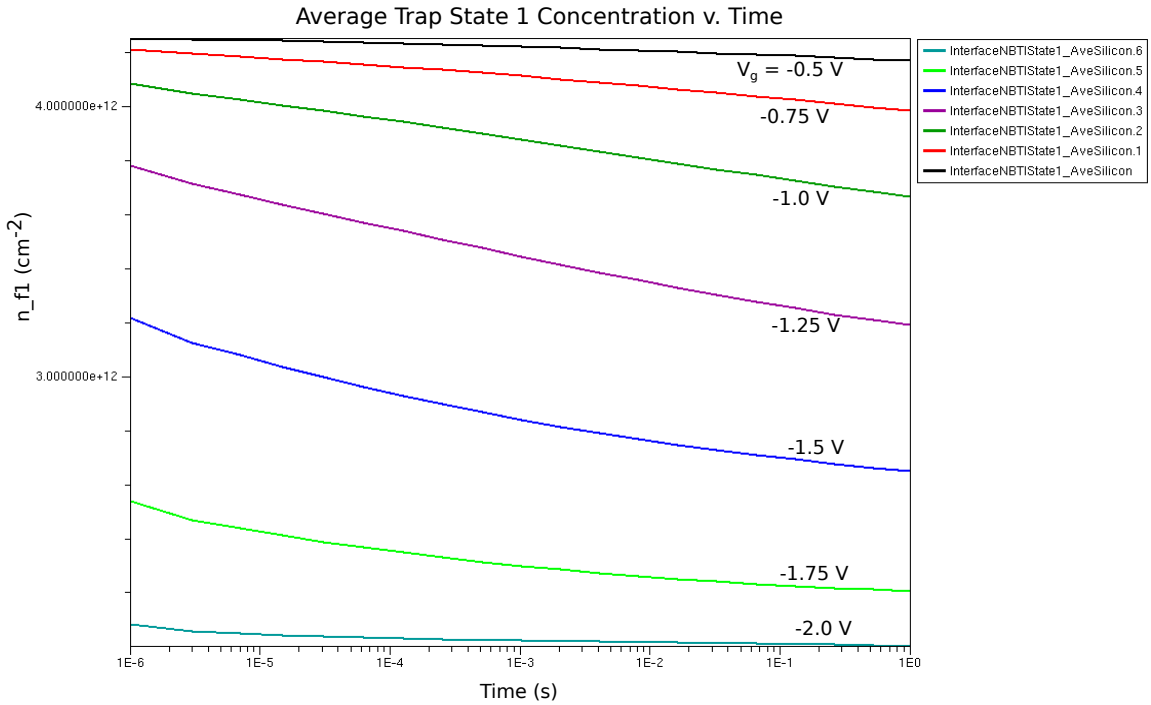


Figure 3.78: Gate voltage variation of $\langle f_1 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K

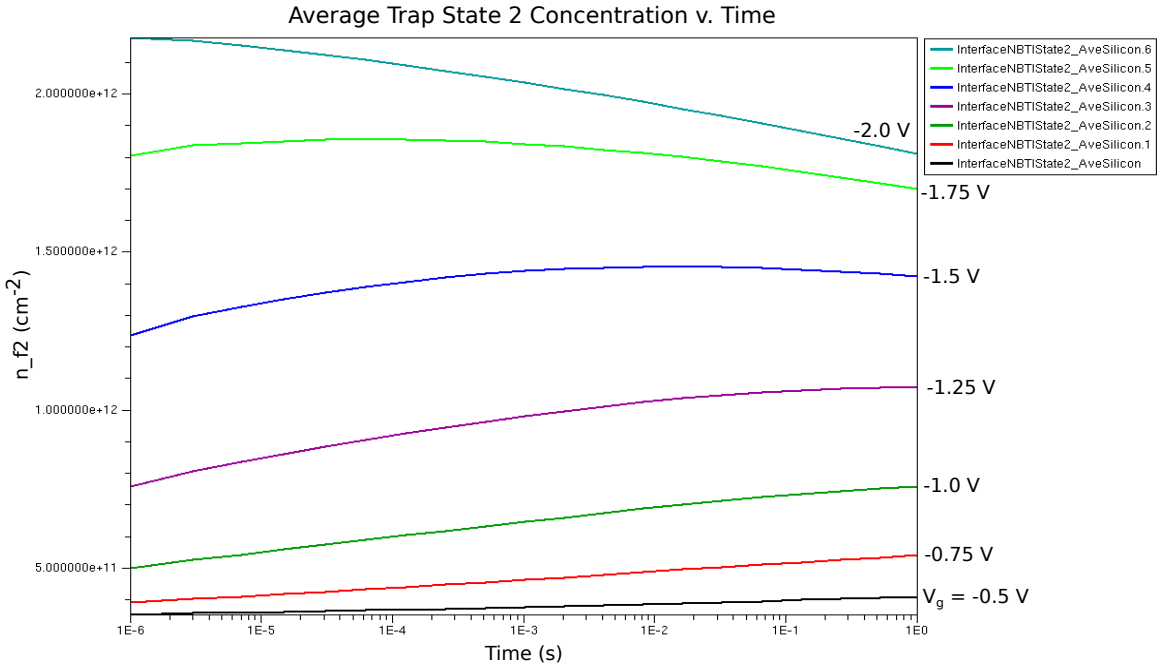


Figure 3.79: Gate voltage variation of $\langle f_2 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K

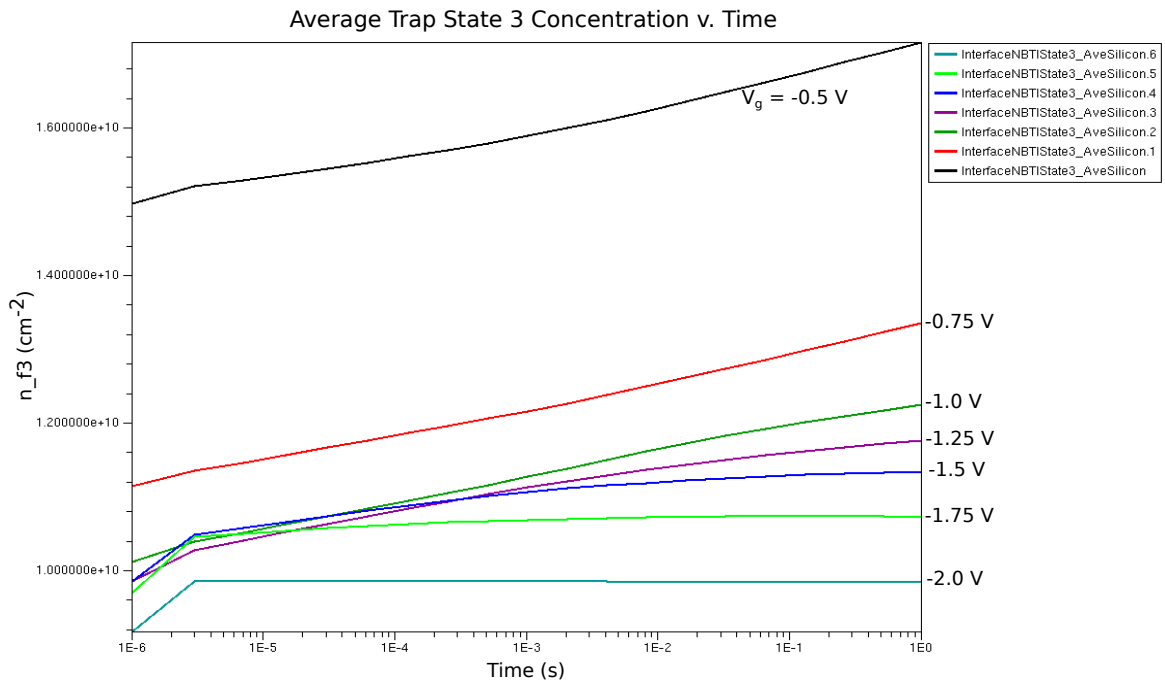


Figure 3.80: Gate voltage variation of $\langle f_3 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K

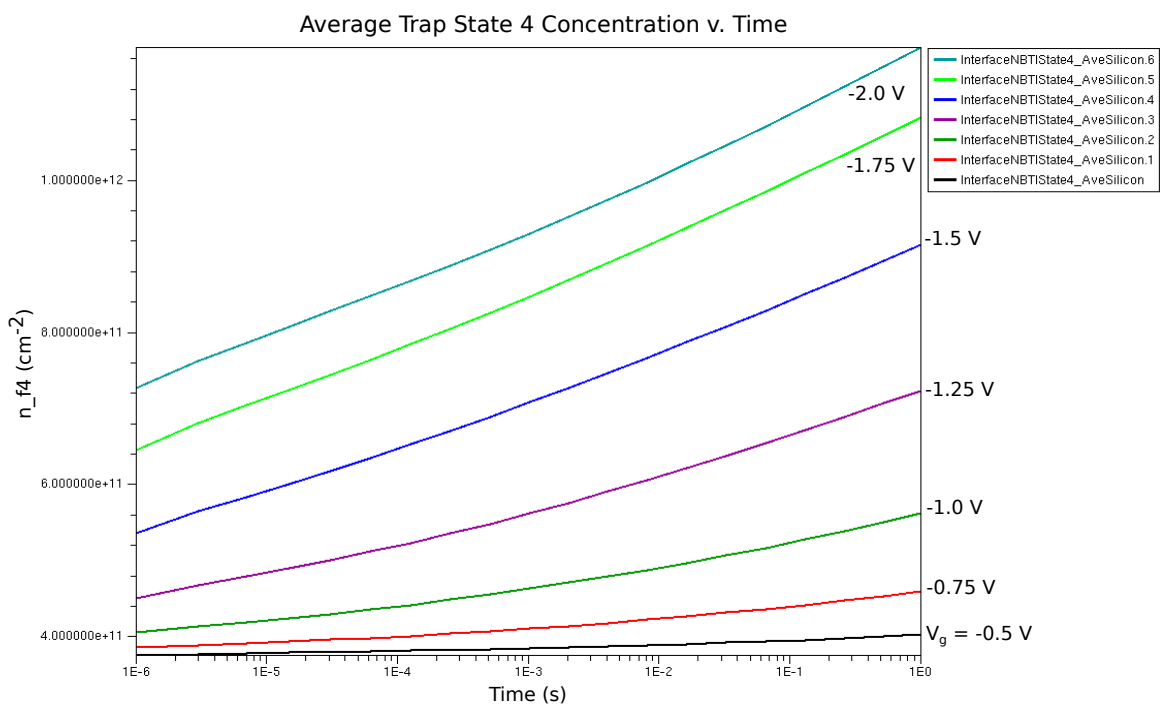


Figure 3.81: Gate voltage variation of $\langle f_4 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K

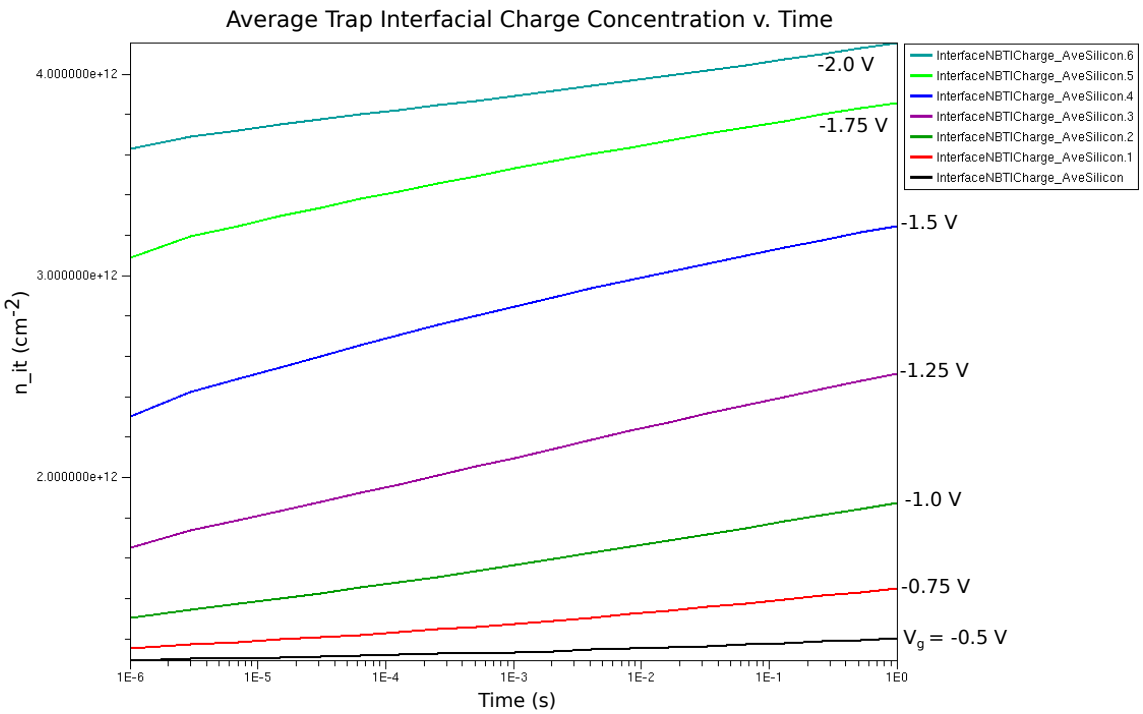
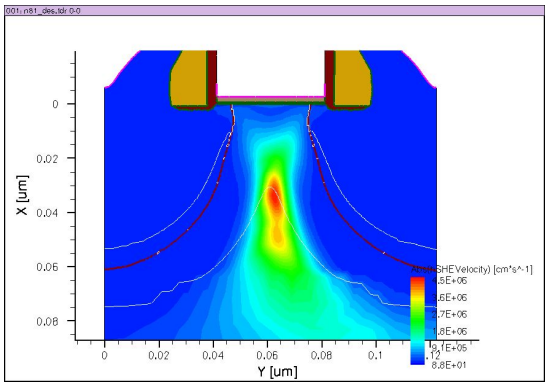


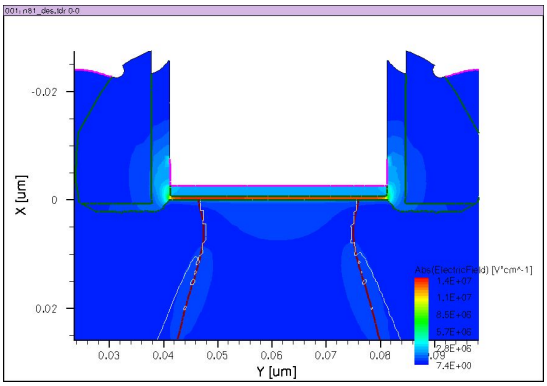
Figure 3.82: Gate voltage variation of Q during symmetric stress, $T_{\text{ext}} = 375$ K

to the 2D FinFET. The structure has a gate length of $L_g = 15$ nm and is given in Figure 3.84(a), where we display the oxynitride spacers and the gate oxide surrounding the channel translucently. The finite element mesh for this structure is shown in Figure 3.84 and was generated using the Sentaraus Delaunay mesh routine. We consider both symmetric and asymmetric stress configurations for a couple gate voltages and an ambient temperature of $T_{\text{ext}} = 375$ K. Due to the additional complexity entailed by the three dimensions of the structure, we are limited to using the hydrodynamic model to simulate hole energy transport. In order to reduce computational cost, we have taken advantage of the symmetry of the device structure and have only simulated half of the full device. This is reflected in the TecPlot images displayed below.

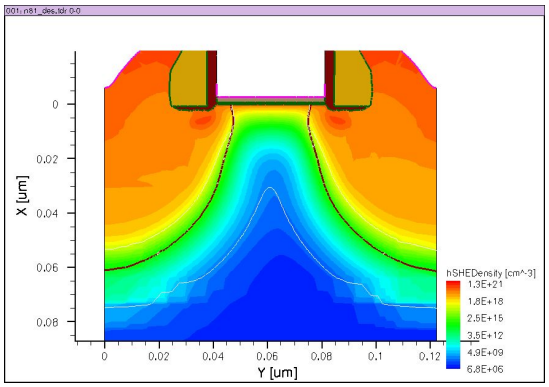
The degradation in threshold voltage for a couple symmetric and asymmetric stress cases are displayed below (Figure 3.85, Figure 3.86, Figure 3.87). We find that



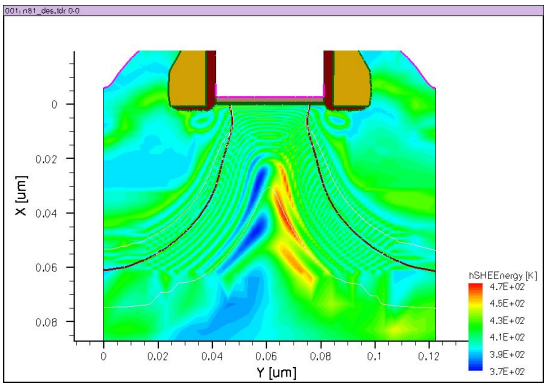
(a) Hole velocity $|\boldsymbol{v}_{\mathrm{h}}|$



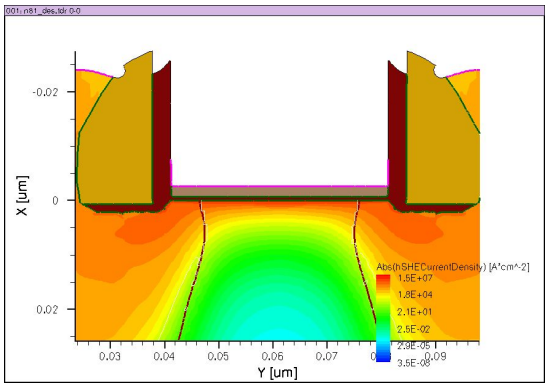
(b) Electric field $|\boldsymbol{E}|$



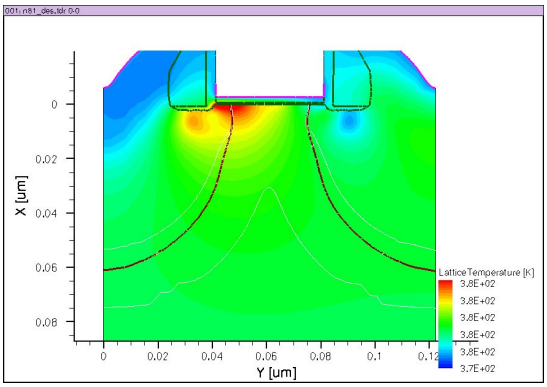
(c) Hole density n_{h}



(d) Hole energy $|\epsilon_{\mathrm{h}}|$

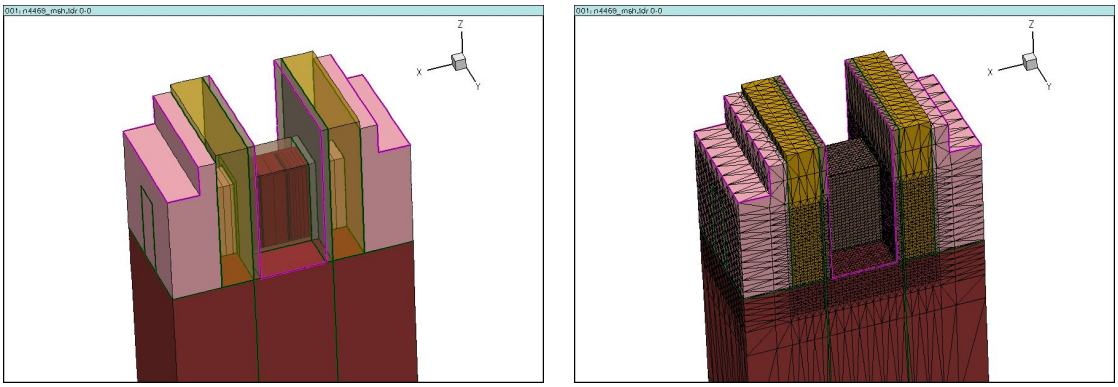


(e) Hole current density $|\boldsymbol{J}_{\mathrm{h}}|$



(f) Lattice temperature T_{L}

Figure 3.83: 2D Planar MOSFET under symmetric stress ($V_{\mathrm{g}} = -1.5$ V, $T_{\mathrm{ext}} = 375$ K)



(a) 3D FinFET structure (b) Finite element mesh

Figure 3.84: 3D FinFET structure with finite element mesh

the shift in V_{th} is lowered compared to the 2D FinFET case, and successive ΔV_{th} are much closer together. Furthermore, the rate of V_{th} degradation again exhibits a break from the expected monotonic behavior for increasing V_{g}. The monotonic behavior of ΔV_{th} is again restored for the asymmetric stress configuration, where again we find that the range of V_{th} degradation is much larger compared to the symmetric stress configuration.

Figure 3.88 to Figure 3.91 show the occupation of the trap states 1-4 under symmetric stress. Again, the time domain of interest for V_{th} degradation is $10^{-3} \leq t \leq 1$ s. For states 1, 2 and 4, we observe normal decrease or increase, respectively, in the occupation number for increasing V_{g}, except for the $V_{\mathrm{g}} = -1.0$ V curve, which consistently follows the -0.5 V curve. Furthermore, we observe an interesting separation in the trap state 3 curve behavior with respect to increasing V_{g}, in which the occupation number initially decreases up to -1.0 V and then increases beginning at -1.25 V. The collective behavior of the trap state curves for $V_{\mathrm{g}} = -1.0$ V indicates that the relaxation of the lattice for trap state 3 dominates over the hole trapping and interface state generation. This behavior is again reflected in the dynamics of the total trapped charge (Figure 3.92).

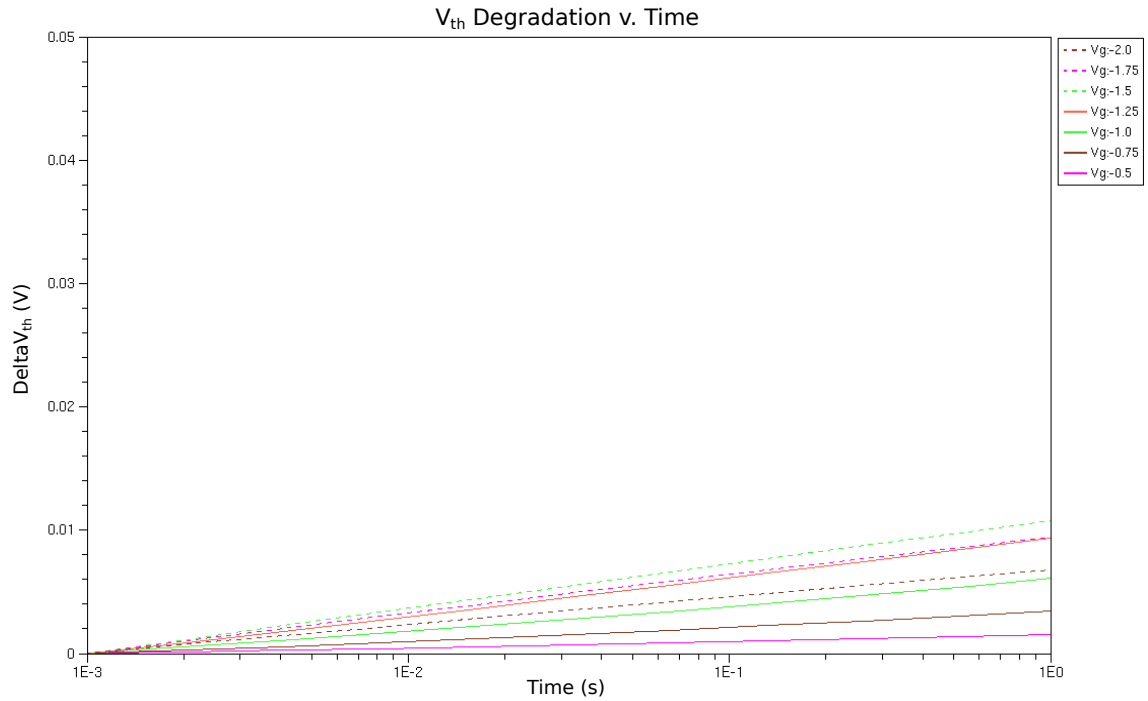


Figure 3.85: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 325$ K

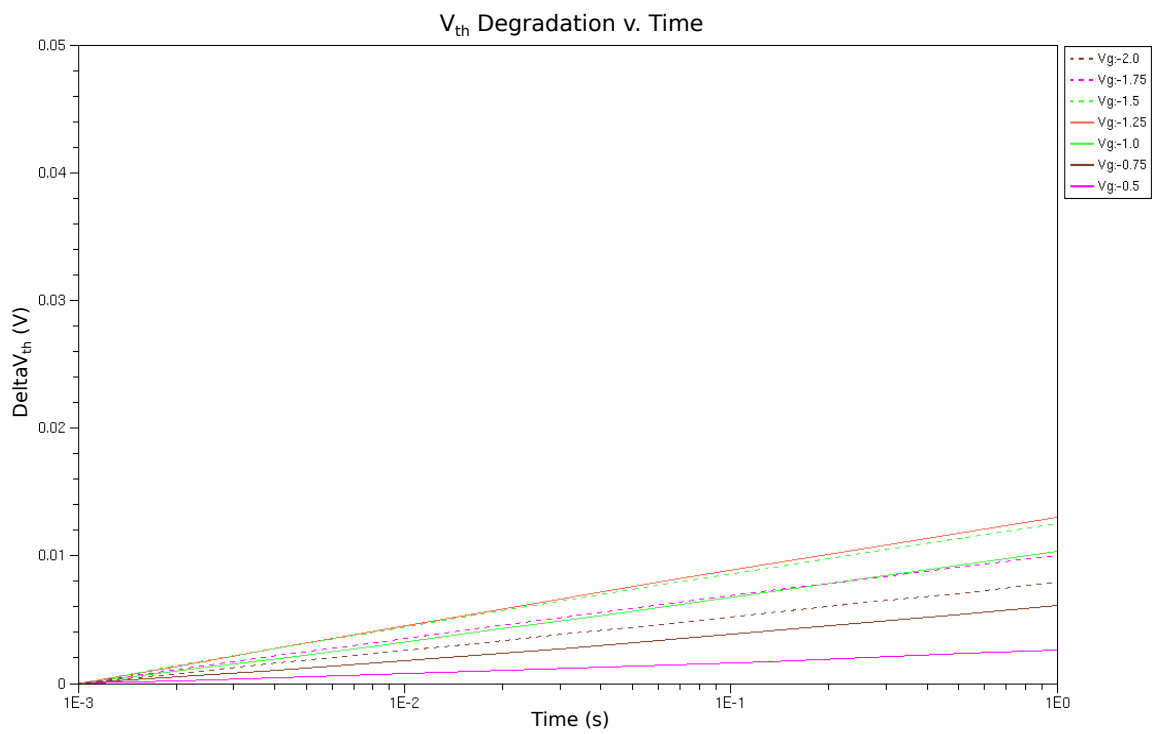


Figure 3.86: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during symmetric stress, $T_{\text{ext}} = 375$ K

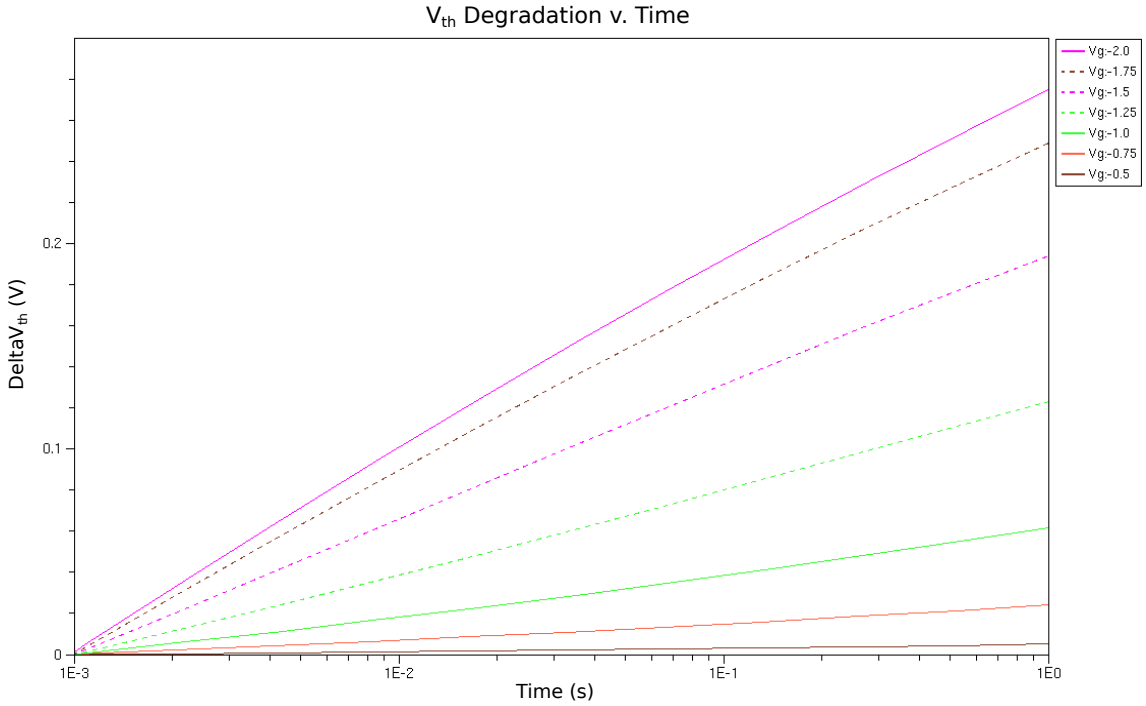


Figure 3.87: Gate voltage variation of $\Delta V_{\text{th}}(t)$ during asymmetric stress ($V_{\text{g}} = V_d$), $T_{\text{ext}} = 375$ K

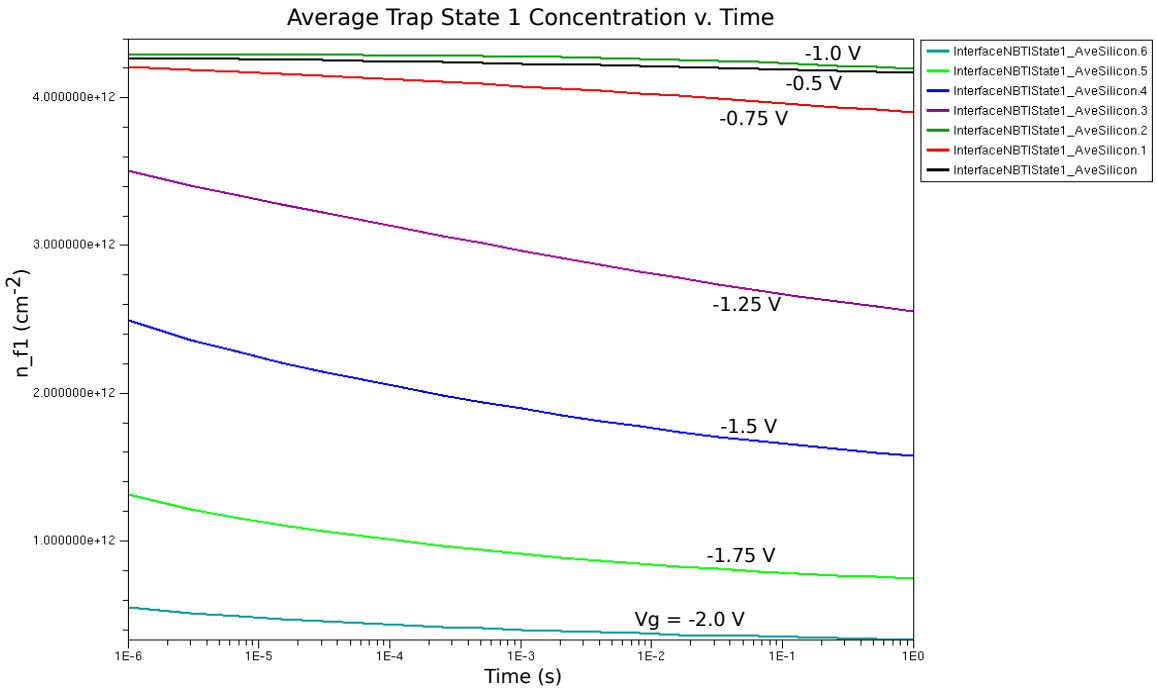


Figure 3.88: Gate voltage variation of $\langle f_1 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K

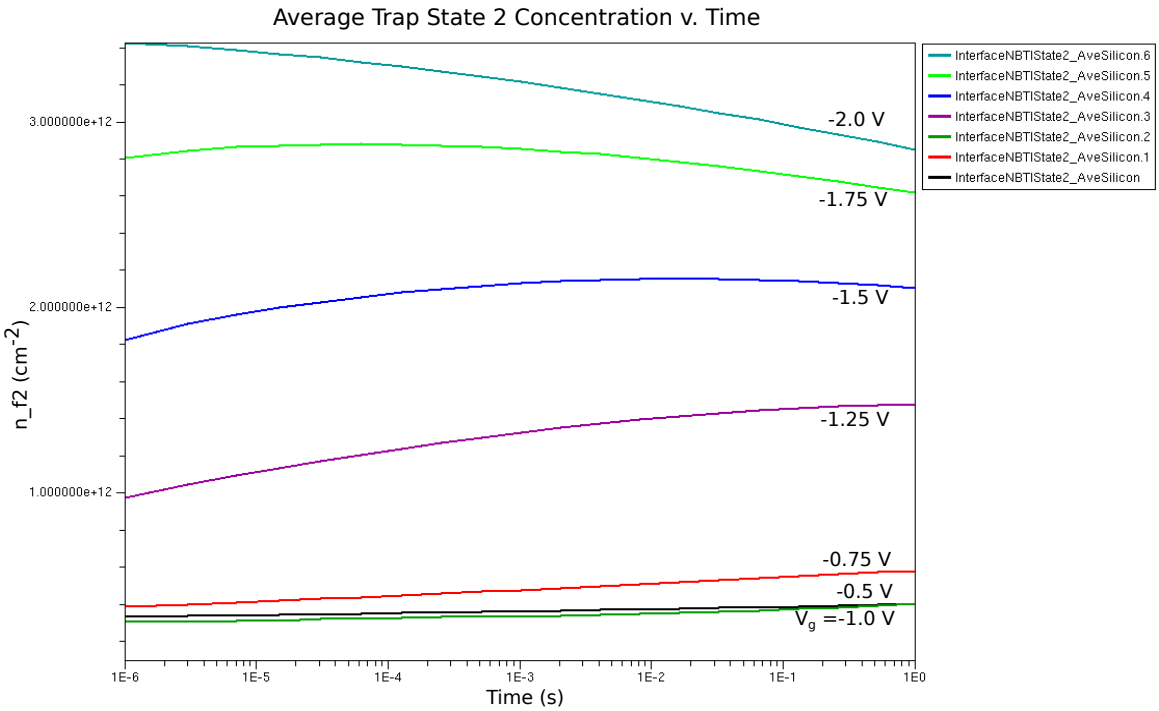


Figure 3.89: Gate voltage variation of $\langle f_2 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K

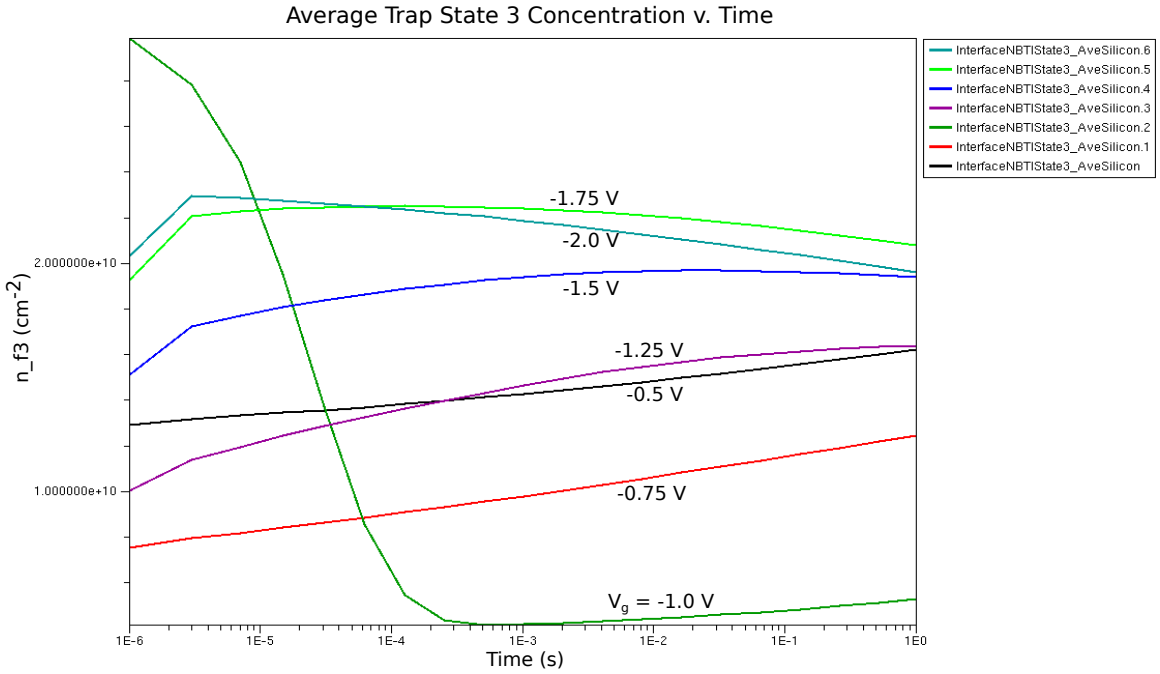


Figure 3.90: Gate voltage variation of $\langle f_3 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K

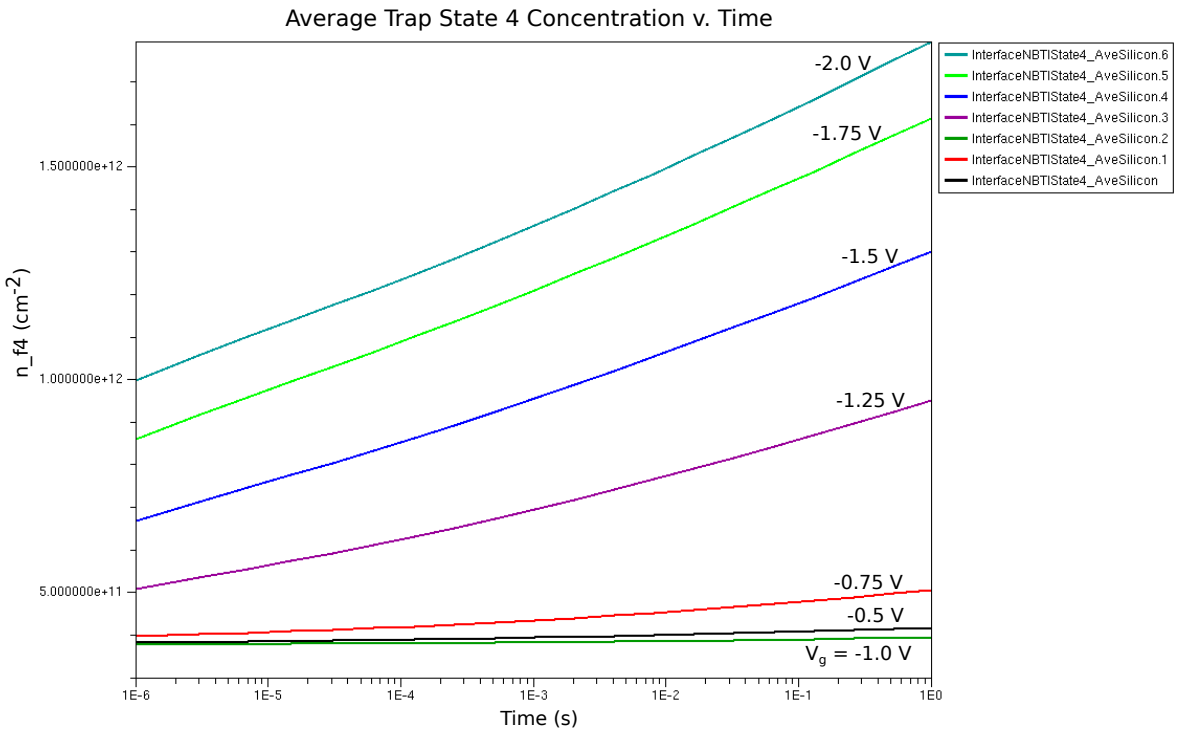


Figure 3.91: Gate voltage variation of $\langle f_4 \rangle$ during symmetric stress, $T_{\text{ext}} = 375$ K

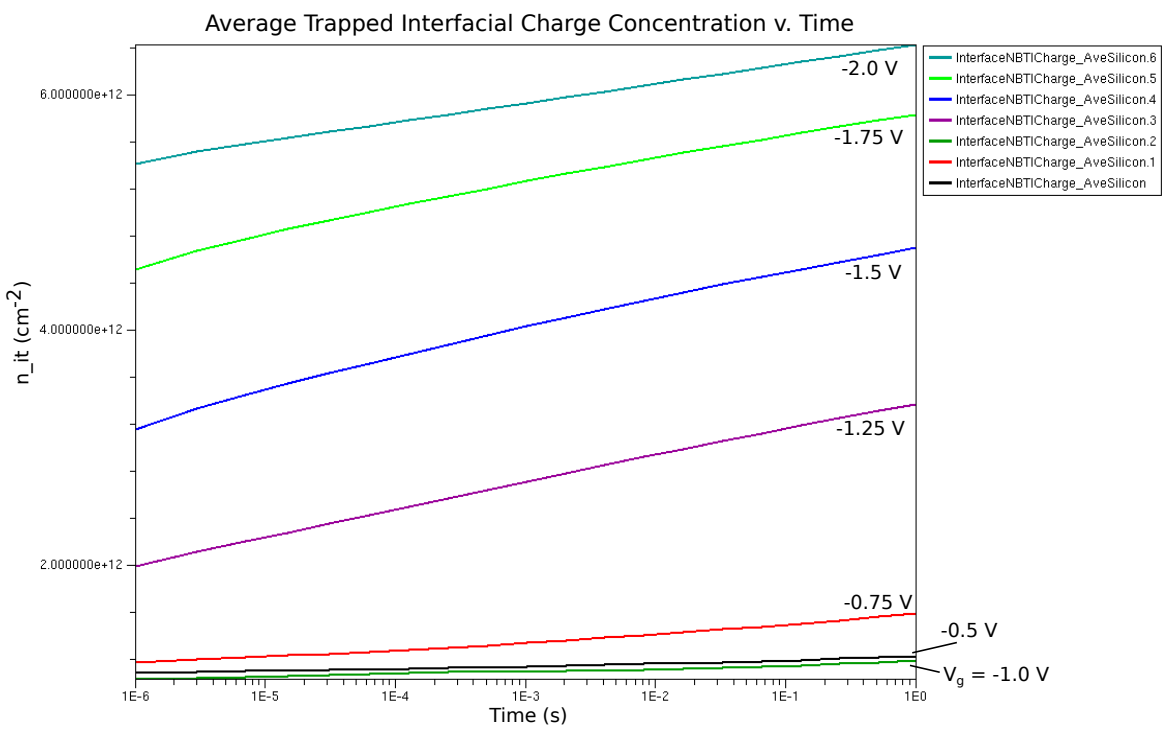


Figure 3.92: Gate voltage variation of Q during symmetric stress, $T_{\text{ext}} = 375$ K

Figure 3.93 to Figure 3.96 show the occupation of the trap states 1-4 under asymmetric stress. The $V_g = -1.0$ V curves exhibit the same behavior as in the symmetric stress configuration, while we find that the -2.0 V curves exhibit a complex crossing behavior for each trap state not observed in the symmetric stress configuration. This is further reflected in the total trapped charge Q shown in Figure 3.97. Comparing the -2.0 V curves for states 2 and 3 to those of states 1 and 4 shows that hole trapping and detrapping begins to dominate over interface state generation and lattice relaxation.

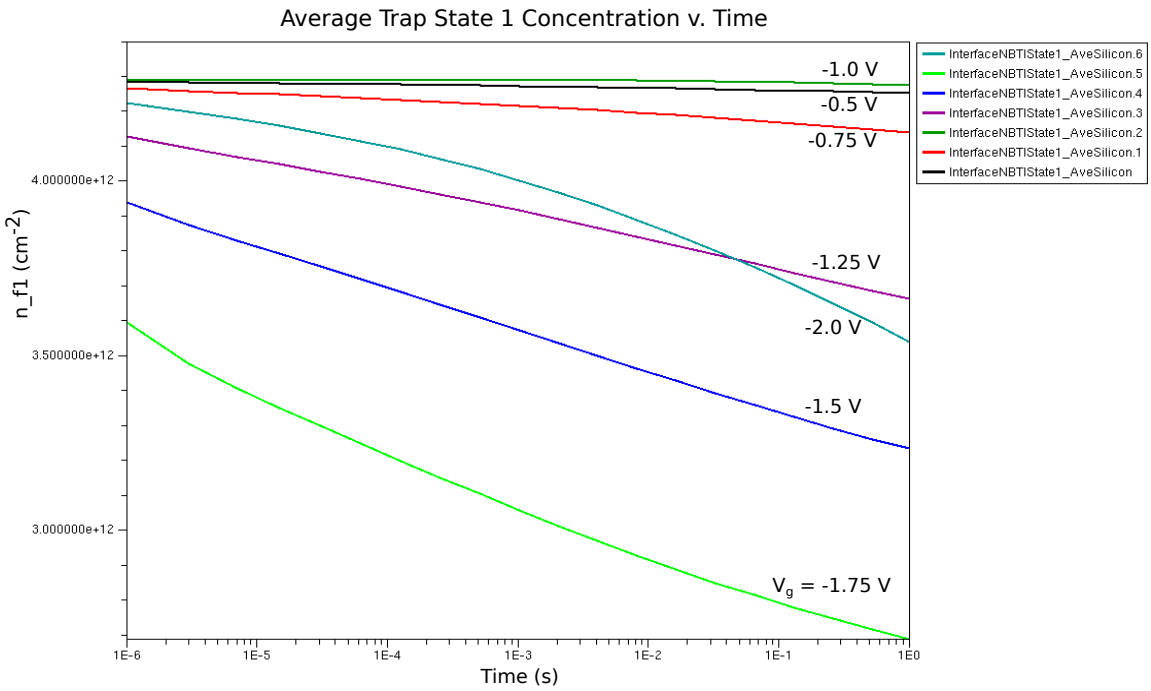


Figure 3.93: Gate voltage variation of $\langle f_1 \rangle$ during asymmetric stress, $T_{\text{ext}} = 375$ K

Below we display the spatial distributions for the hole velocity, electric field, hole density, hole temperature, hole current density and lattice temperature for two gate voltages, -1.0 and -1.5 V, under symmetric stress. For clarity, we give two views of the distribution. The cross-sectional view shows the distribution across a vertical plane through the channel dividing the device in half. The outer-channel view shows the distribution on the outer plane of the channel where we have removed the gate

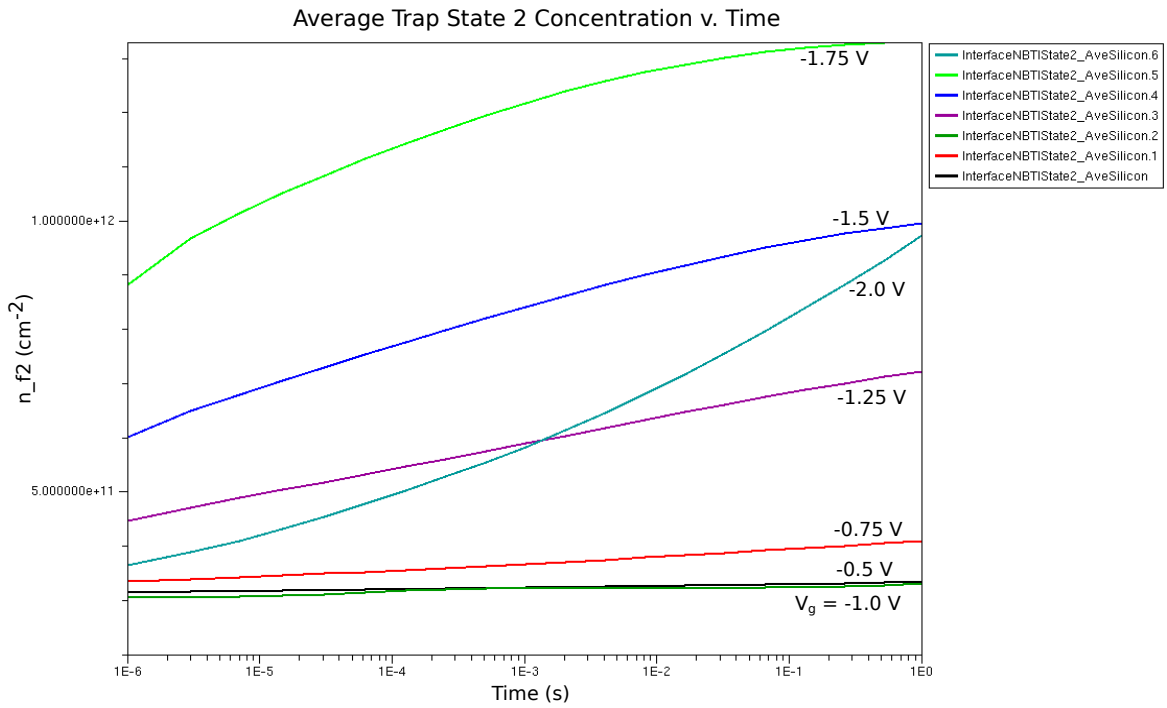


Figure 3.94: Gate voltage variation of $\langle f_2 \rangle$ during asymmetric stress, $T_{\text{ext}} = 375$ K

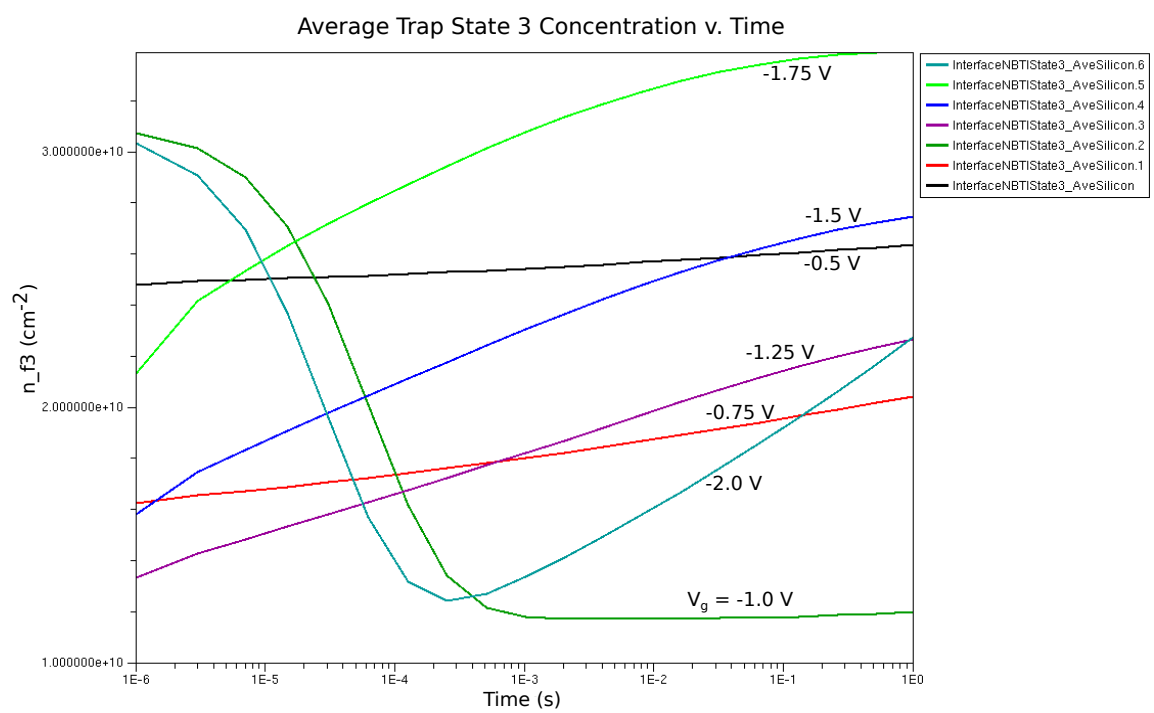


Figure 3.95: Gate voltage variation of $\langle f_3 \rangle$ during asymmetric stress, $T_{\text{ext}} = 375$ K

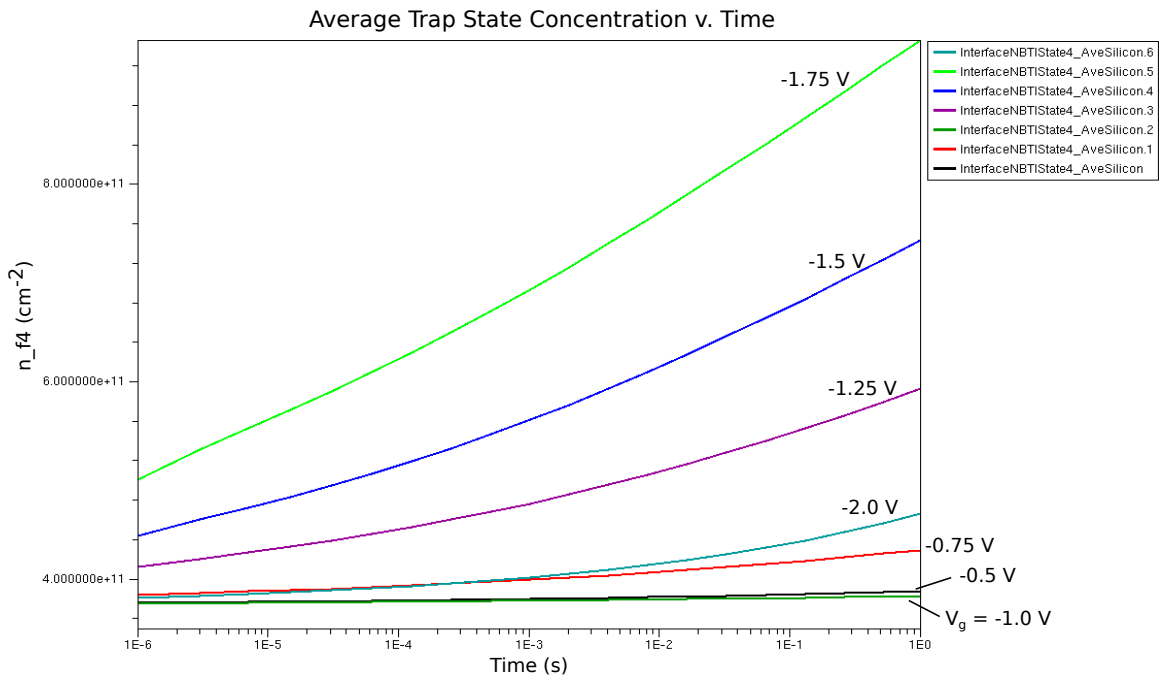


Figure 3.96: Gate voltage variation of $\langle f_4 \rangle$ during asymmetric stress, $T_{\text{ext}} = 375$ K

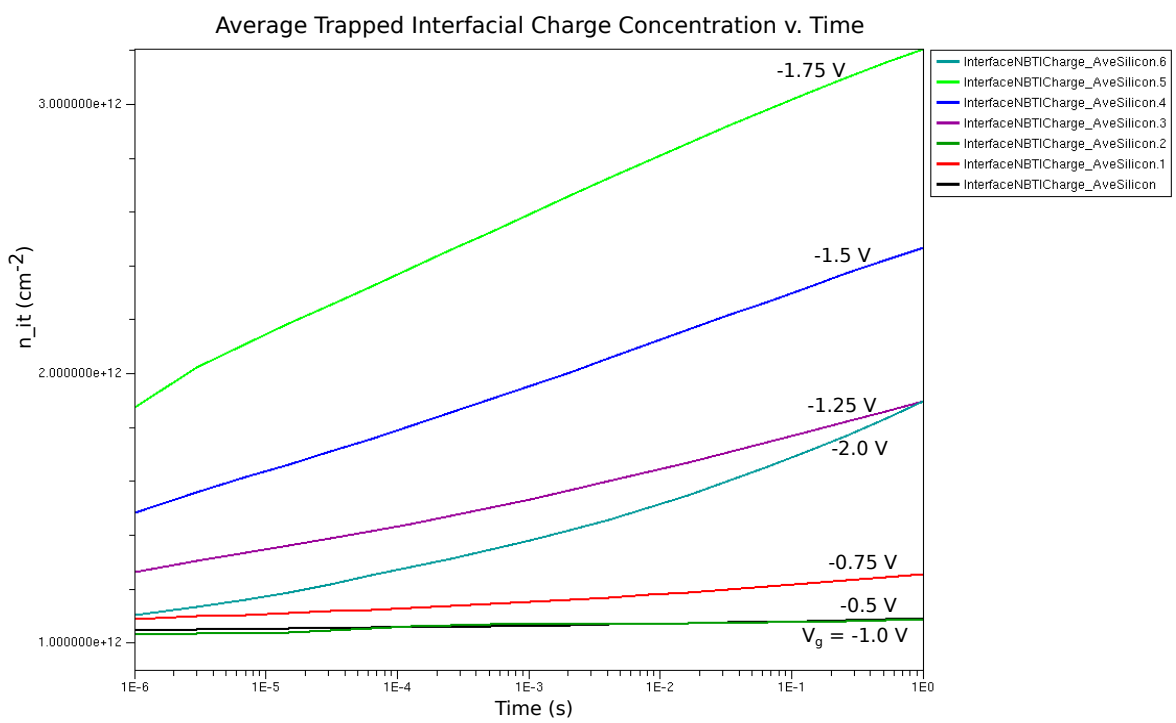
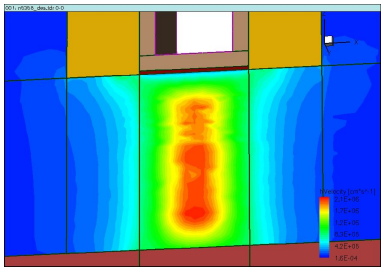


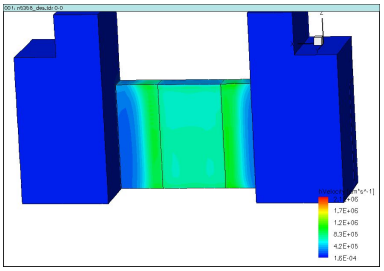
Figure 3.97: Gate voltage variation of Q during asymmetric stress, $T_{\text{ext}} = 375$ K

oxide. The orientation of the device in the outer-channel view is rotated around the vertical axis of the device compared to the cross-sectional view.

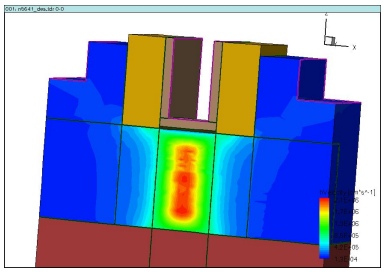
Figure 3.98 shows the distribution of hole velocity in the channel. We find that the maximum is localized in the center of the channel, similar to the 2D FinFET, while the spread of the maximum region decreases across the width of the channel for higher V_g. In Figure 3.99, we find that the electric field is again highly localized near the gate, through the gate dielectric. We observe in Figure 3.100 that the hole density is minimal in the center of the channel under the gate, while the outer-channel views show an increase in hole density near the corners of the channel for increasing V_g



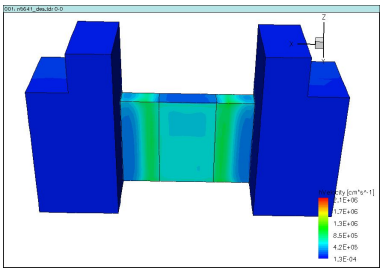
(a) $V_g = -1.0$ V - Cross-sectional view



(b) $V_g = -1.0$ V - Outer-channel view



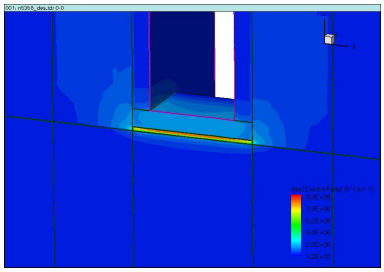
(c) $V_g = -1.5$ V - Cross-sectional view



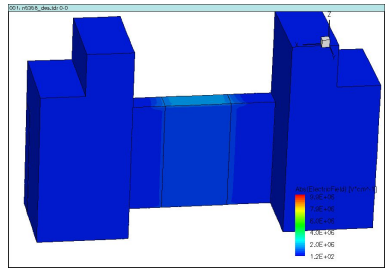
(d) $V_g = -1.5$ V - Outer-channel view

Figure 3.98: Hole velocity $|\boldsymbol{v}_h|$ under symmetric stress ($T_{ext} = 375$ K)

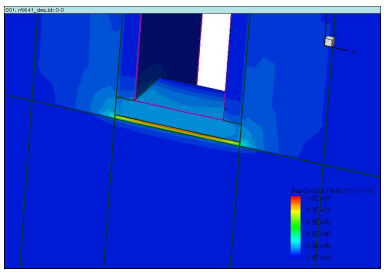
The maximum and minimum regions of hole temperature (which is the hydrodynamic equivalent to the hole energy given in the SHE model) spread across the width and height of the channel for higher V_g, as shown in Figure 3.101. Furthermore, we observe a second temperature maximum develop in the channel near the source-side



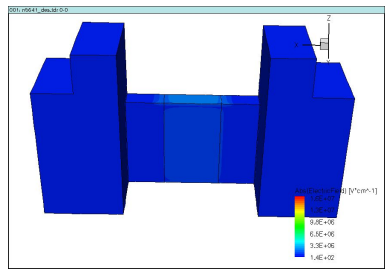
(a) $V_{\mathrm{g}} = -1.0$ V - Cross-sectional view



(b) $V_{\mathrm{g}} = -1.0$ V - Outer-channel view

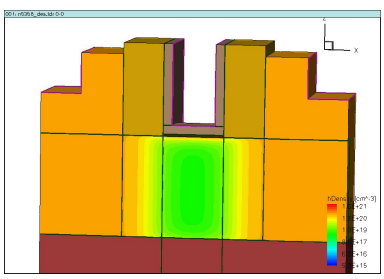


(c) $V_{\mathrm{g}} = -1.5$ V - Cross-sectional view

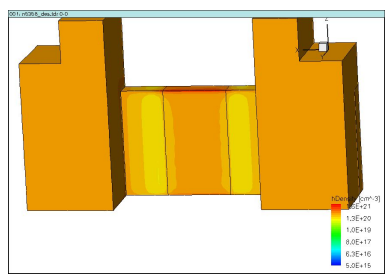


(d) $V_{\mathrm{g}} = -1.5$ V - Outer-channel view

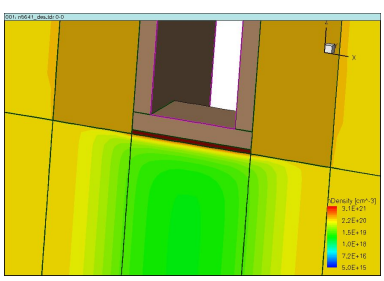
Figure 3.99: Electric field $|\boldsymbol{E}|$ under symmetric stress ($T_{\mathrm{ext}} = 375$ K)



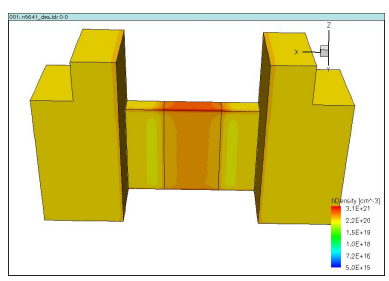
(a) $V_{\mathrm{g}} = -1.0$ V - Cross-sectional view



(b) $V_{\mathrm{g}} = -1.0$ V - Outer-channel view



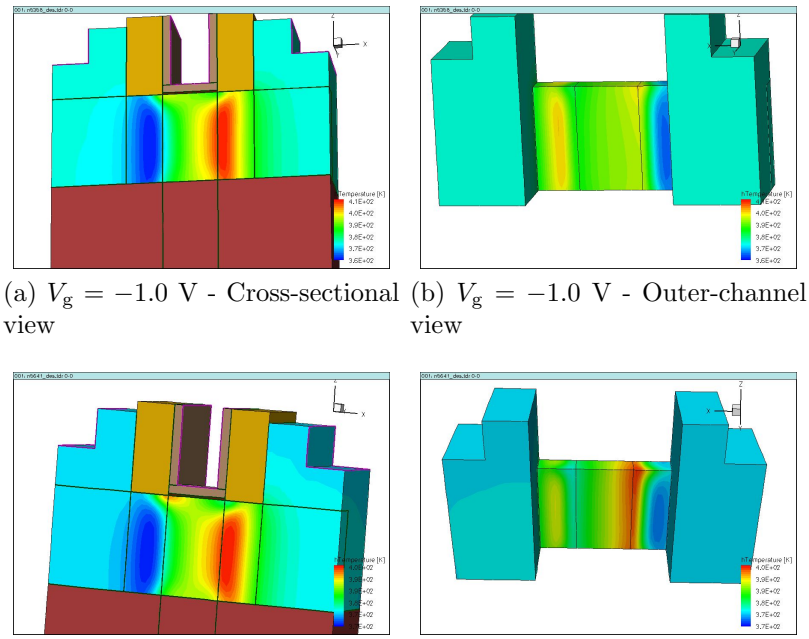
(c) $V_{\mathrm{g}} = -1.5$ V - Cross-sectional view



(d) $V_{\mathrm{g}} = -1.5$ V - Outer-channel view

Figure 3.100: Hole density n_{h} under symmetric stress ($T_{\mathrm{ext}} = 375$ K)

corner of the gate for $V_g = -1.5$ V (Figure 3.101(c)). In Figure 3.102, we find that the hole current density is greatest at the channel surfaces beneath the gate and increases for increasing V_g.



(a) $V_g = -1.0$ V - Cross-sectional view

(b) $V_g = -1.0$ V - Outer-channel view

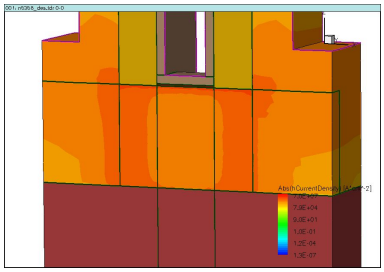
(c) $V_g = -1.5$ V - Cross-sectional view

(d) $V_g = -1.5$ V - Outer-channel view

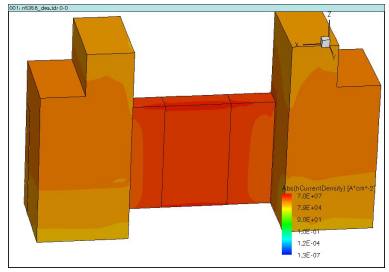
Figure 3.101: Hole temperature T_h under symmetric stress ($T_{ext} = 375$ K)

The lattice temperature under symmetric stress and Dirichlet thermal boundary conditions develops a small hot spot on the order of 10 K above the boundary temperature in the drain side of the channel for both $V_g = -1.0$ and -1.5 V (Figure 3.103). This hot spot spreads further across the width of the channel for higher V_g. Furthermore, we find that the center of the hot spot is closer to the SOI substrate than the top of the gate, which follows from the relative proximities of these regions to the thermally conductive boundaries.

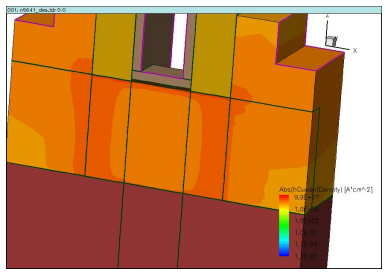
If we consider the qualitative similarities between the hydrodynamic and SHE solutions for the hole velocity and hole current density distributions, in particular that the maximum in the hole current density does not extend across the channel in



(a) $V_g = -1.0$ V - Cross-sectional view

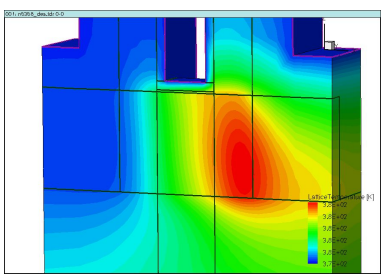


(b) $V_g = -1.0$ V - Outer-channel view

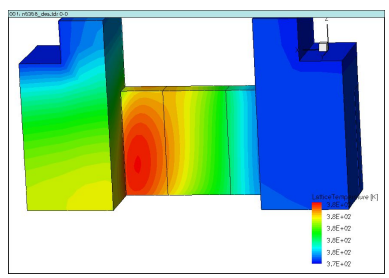


(c) $V_g = -1.5$ V - Cross-sectional view

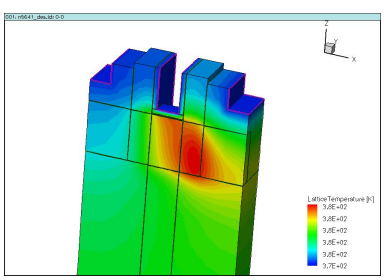
Figure 3.102: Hole current density $|\boldsymbol{J}_h|$ under symmetric stress ($T_{ext} = 375$ K)



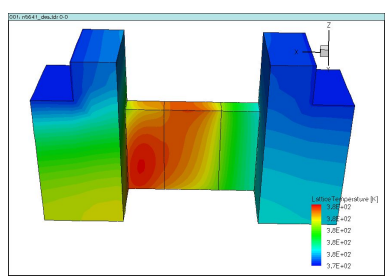
(a) $V_g = -1.0$ V - Cross-sectional view



(b) $V_g = -1.0$ V - Outer-channel view



(c) $V_g = -1.5$ V - Cross-sectional view



(d) $V_g = -1.5$ V - Outer-channel view

Figure 3.103: Lattice temperature T_L under symmetric stress ($T_{ext} = 375$ K)

the asymmetric case as in the symmetric case, then we can take the 2D FinFET to be a baseline for our analysis of the breaking of the monotonicity in the V_{th} degradation rate under symmetric stress for the 3D FinFET. While in the 3D FinFET case, we find an additional contribution to this ΔV_{th} behavior due to the dynamics of the trap states at different V_{g}, the nature of the hole velocity and the hole current density may also play a contributing role. This points the way to potential future work.

CHAPTER 4

CONCLUSIONS

4.1 Research Results and Outlook

NBTI is expected to become an even greater reliability concern in deeply scaled pMOSFETs. This is not only due to the implementation of new materials within the transistor structure, but also to the presence of higher electric fields and more significant heat generation in confined geometries [7, 41]. Being able to predict NBTI degradation is crucial for the development and long-term success of novel transistor structures, such as the FinFET. Thus, this work constitutes a first step toward predictive NBTI modeling in FinFET structures, including contributions from self-heating. In particular, we have chosen a two-stage NBTI model that allows us to readily incorporate the electric field and temperature dependence as well as the recoverable component of NBTI degradation. This two-stage NBTI model hinges on the dynamics of deep-level defects (the E′- and P_b-centers) in SiO_2 and at Si/SiO_2 interfaces.

Furthermore, we have provided a detailed overview of the theory underlying charge transport simulation, self-heating phenomena, and the nonradiative capture of charge by multiphonon emission that leads to two-stage NBTI. Through this discussion, we have attempted to demonstrate the trade-offs that must be made between model accuracy and computational efficiency in simulating complex semiconductor structures. This work is characterized by the implementation of a hierarchy of approximations, necessary to achieve proper convergence of our TCAD finite element device simulation. This hierarchy is apparent in the derivation of the hydrodynamic equations from the Boltzmann transport equation and the high-temperature limit of the capture cross-sections for deep-level nonradiative transitions.

In this work, then, we have analyzed the qualitative behavior of V_{th} shift for short stress times due to NBTI degradation and self-heating by self-consistently solving a coupled set of energy transport, temperature, and NBTI trap state equations with TCAD simulation. We have considered two stress configurations and three device structures: a 2D FinFET, a 2D MOSFET, and a 3D FinFET. Furthermore, we have separately employed a spherical harmonic expansion model for the 2D cases and a hydrodynamic model for the 3D FinFET, while comparing the differences in the two for the 2D FinFET. We have also briefly considered the effect of thermal boundary conditions on our results.

For the 2D FinFET, we observe a logarithmic time dependence for the shift in the threshold voltage. Furthermore, we find that while V_{th} degradation is worsened for increasing boundary temperatures, an improvement in this degradation occurs past a certain bias for increasing V_{g} when the device is under symmetric stress and at lower temperatures. This is contrary to the typically observed behavior of V_{th} shift due to NBTI. In this case, we find that the 2D FinFET does not exhibit a monotonic increase in the rate of V_{th} degradation, and we have attributed this breaking of monotonic behavior to the complex interaction between the trap states and the degradation of the drain current due to mechanisms independent from NBTI but activated during NBTI stress. Additionally, we find that this monotonic behavior is restored for higher boundary temperatures and for asymmetric stress configurations. In particular, we have noted the potential contribution of velocity overshoot to drain current degradation, as well as the effects of the hole current density distribution in the device channel. The increase in lattice temperature during asymmetric stress may also contribute to the degradation in the drain current. We have also observed a similar ΔV_{th} behavior in the 2D MOSFET under symmetric stress, which indicates that the presence of velocity overshoot and the distribution of the current density may be intrinsically overestimated by the SHE and hydrodynamic models.

For the 3D FinFET, we have likewise observed a similar breaking of monotonic behavior in the V_{th} shift. We also find a strikingly different behavior of the NBTI trap states, which to some degree contributes to degradation of the drain current. Indeed, this case exhibits an intricate interaction between hole trapping and detrapping, interface state formation, and lattice relaxation. Nonetheless, we expect similar complications arising from velocity overshoot and the hole current density distribution.

For both the 2D and 3D FinFETs, the self-heating of the lattice due to hole scattering is insignificant under symmetric stress with Dirichlet thermal boundary conditions. Under asymmetric stress, we observe the emergence of a much more pronounced hot spot near the drain. However, the self-heating model employed in these cases does not treat momentum and energy transfer between the coupled hole, optical phonon, and acoustic phonon systems. We thus expect the emergence of a hot spot with much higher temperature due to the phonon bottleneck, which would require a more accurate model to simulate. Nonetheless, this hot spot likely contributes to the degradation in drain current during asymmetric stress.

Simulation convergence and computational requirements severely hinder the degree to which we can account for the underlying physics of heat generation and NBTI degradation. This problem is made more immediate by the increasingly smaller scales of the transistor. For a FinFET with 15 nm gate length, we should expect the discrete nature of the lattice and the wave nature of the charge carriers to play an increasingly important role in transport and trapping phenomena. Indeed, it is very likely that the non-monotonicity observed in the V_{th} shift will be alleviated by a more fundamental model, by correcting for such phenomena as velocity overshoot and underestimated hot spot temperatures. This is crucial if predictive reliability simulations are to be accurate. An alternative to implementing new models is to calibrate the existing ones against experimental data, although one then sacrifices the predictive nature of

the simulation. In fact, it could be that calibrating the simulation parameters employed in this work could correct for the non-monotonic behavior in V_{th} shift. Since this phenomena has not been experimentally observed in pMOSFETs, we may find that calibration will push the conditions under which the non-monotonic phenomena emerges outside the normal range of typical experiments.

4.2 Recommendations for Future Work

The results of this work point significantly to the need for model calibration against experimental data and for the implementation of more accurate physical models in device simulation [41]. Approximate models such as the hydrodynamic and SHE models are quickly losing their physical validity for aggressively scaled semiconductor devices such as the FinFET, and the problem will only become more apparent for smaller device architectures. On these length scales, one should expect to approach the limit to which these models can be empirically calibrated [16, 17]. Thus, we must look toward new models to account for nanoscale phenomena.

A first step toward correcting the apparent velocity overshoot would be to include more higher order moments of the BTE in the set of coupled device equations being solved by finite elements. The six moments model has demonstrated this gain in physical accuracy while still maintaining computational efficiency [63]. This model should be immediately compatible with the two-stage NBTI model. However, it effectively ignores the existence of different phonon modes and the nonequilbrium states that can develop between them.

Direct solutions to the electron or hole BTE coupled with the energy balance equations for acoustic and optical phonon modes or with the full phonon BTE can be tractable using Monte Carlo methods in 2D [52, 53, 54, 55]. Open source codes, such as MONET, are particularly adept at treating phonon scattering processes during charge transport using a frozen-field approximation. However, implementing such Monte Carlo methods in NBTI simulation would require one to generalize the two-stage

NBTI model to couple with acoustic and optical phonon modes, or with an effective phonon heat bath. Furthermore, obtaining real transient solutions to the charge carrier and phonon distribution functions requires one to account for the mutual interaction between the NBTI trap states and the charge carrier and phonon systems. Thus, one would be required to self-consistently iterate the numerical solver over the time evolution of the system, which may be too computationally expensive.

To account for quantum mechanical effects, one often employs a quantum correction in the form of an additional potential in the macroscopic device equations, or a density matrix method [29, 70]. Currently, the two-stage NBTI model is not compatible with such quantum mechanical corrections. Nonequilbirium Green's function (NEGF) methods readily account for quantum mechanical effects as well as particle interactions [65]. Thus, one could in principle model quantum transport through a nanoscale device including hole-phonon interactions using NEGF methods [64, 71, 72, 73, 74] . This approach has been implemented in modeling carbon nanotubes and molecular systems [67, 68]. The NEGF approach is unfeasible for a three-dimensional device such as the FinFET because the number of nodes required would make the computation time intractable. However, the implementation of NEGF methods for one or two dimensions of the device could prove valuable [69, 76]. One would again be faced with the problem of generalizing the two-stage NBTI model.

REFERENCES CITED

[1] H.-S. P. Wong, "Beyond the Conventional Transistor," *IBM J. Res. and Dev.*, vol. 46, no. 2/3, pp. 133 - 168, March/May 2002.

[2] J-P. Colinge, *FinFETs and Other Multigate Transistors* (Springer, New York, 2008).

[3] J. Mar, "The Application of TCAD in Industry," *International Conference on Simulation of Semiconductor Processes and Device*, pp. 139 - 145, September 1996.

[4] A. Jackson, "Cyber-Enabled Discovery and Innovation," *AMS Notices*, vol. 54, no. 6, pp. 752 - 754, June/July 2007.

[5] D. Hisamoto, W.-C. Lee, *et al.*, "FinFET - A Self-Aligned Double-Gate MOSFET Scalable to 20 nm," *IEEE Trans. Electron Devices*, vol. 47, no. 12, pp. 2320 - 2325, December 2000.

[6] A. Strong, *et al*, "Introduction," in *Reliability Wearout Mechanisms in Advanced CMOS Technologies* (IEEE Press Series on Microelectronic Systems, Wiley, New Jersey, 2009).

[7] G. LaRosa, "Negative Bias Temperature Instability in pMOSFET Devices," in *Reliability Wearout Mechanisms in Advanced CMOS Technologies* (IEEE Press Series on Microelectronic Systems, Wiley, New Jersey, 2009).

[8] M. A. Stroscio, M. Dutta, *Phonons in Nanostructures* (Cambridge University Press, UK, 2001).

[9] K. Raleva, D. Vasileska, S. Goodnick, M. Nedjalkov, "Modeling Thermal Effects in Nanodevices," *IEEE Trans. Electron Devices*, vol. 55, no. 6, pp. 1306-1316, June 2008.

[10] D. Vasileska, K. Raleva, S. Goodnick, "Modeling Heating Effects in Nanoscale Devices: The Present and Future," *J. Comput. Electron.*, vol. 7, pp. 66 - 93, July 2008.

[11] E. Pop, S. Sinha, K. Goodson, "Heat Generation and Transport in Nanometer-Scale Transistors," *Proceedings of the IEEE*, vol. 94, no. 8, pp. 1587 - 1601, August 2006.

[12] S. Sinha, K. E. Goodson, "Review: Multiscale Thermal Modeling in Nanoelectronics," *International Journal for Multiscale Computational Engineering*, vol. 3, no. 1, pp. 107 - 133, 2005.

[13] D. Cahill, W. Ford, *et al.*, "Nanoscale Thermal Transport," *J. App. Phys.*, vol. 93, no. 2, pp. 793 - 818, January 2003

[14] S. Sinha, K. Goodson, "Thermal Conduction in Sub-100 nm Transistors," *Microelectronics Journal*, vol. 37, pp. 1148 - 1157, 2006.

[15] S. Sinha, E. Pop, R. Dutton, K. Goodson, "Non-Equilibrium Phonon Distributions in Sub-100 nm Silicon Transistors," *Transactions of the ASME*, vol. 128, pp. 638 - 647, July 2006.

[16] K. Xiu, "Simulation of Phonon-Limited Mobility for Nano-Scale Si-Based n-Type Non-Planar Device under General Orientation and Stress Condition," *J. Comput. Electron.*, vol. 7, pp. 172 - 175, December 2007.

[17] J. Rowlette, K. Goodson, "Fully Coupled Nonequilibrium Electron-Phonon Transport in Nanometer-Scale Silicon FETs," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 220 - 232, January 2008.

[18] A. J. Scholten, *et al*, "Experimental Assessment of Self-Heating in SOI FinFETs," *IEEE Electron Devices Meeting (IEDM)*, pp. 1 - 4, December 2009.

[19] Y-K. Choi, J-W. Han, H. Lee, "Reliability Issues in Multi-Gate FinFETs," *IEEE Trans. on Solid-State and Integrated Circuit Technology*, pp. 1101 - 1104, 2006.

[20] Y. Wang, *et al*, "Ballistic Phonon Enhanced NBTI," *IEEE International Reliability Physics Symposium*, pp. 258 - 263, 2007.

[21] S. Mahapatra, *et al*, "A Critical Re-evaluation of the Usefulness of the R-D Framework in Predicting NBTI Stress and Recovery," *IEEE International Reliability Physics Symposium*, pp. 6A.3.1 - 6A.3.10, April 2011.

[22] G. D. Mahan, "Quantum Transport Equation for Electric and Magnetic Fields," *Physical Reports - Review Section of Physics Letters*, vol. 145, no. 5, pp. 251 - 318, 1987.

[23] P. Danielewicz, "Quantum Theory of Nonequilibrium Processes I," *Annals of Physics*, vol. 152, pp. 239 - 304, 1984.

[24] T. Grasser, T.-W. Tang, H. Kosina, S. Selberherr, "A Review of Hydrodynamic and Energy-Transport Models for Semiconductor Device Simulation," IEEE Proceedings, vol. 91, no. 2, February 2003.

[25] T. Grasser, *et al*, "Advanced Transport Models for Sub-Micrometer Devices," *Proc. SISPAD*, pp. 1 - 8, 2004.

[26] T. Grasser, H. Kosina, S. Selberherr, "Hot Carrier Effects Within Macroscopic Transport Models," *International Journal of High Speed Electronics and Systems*, vol. 13, no. 3, pp. 873 - 901, 2003.

[27] A. Gnudi, *et al*, "Two-Dimensional MOSFET Simulation by Means of a Multidimensional Spherical Harmonics Expansion of the Boltzmann Transport Equation," *Solid-State Electronics*, vol. 36, no. 4, pp. 575 - 581, 1993.

[28] V. Sverdlov, "Demands of Transport Modeling in Advanced MOSFETs," *Strain-Induced Effects in Advanced MOSFETs* (Springer-Verlag, Berlin, 2011).

[29] V. Sverdlov, *et al*, "Current Transport Models for Nanoscale Semiconductor Devices," *Materials Science and Engineering R*, vol. 58, pp. 228 - 270, 2008.

[30] C. Jacoboni, L. Reggiani, "The Monte Carlo Method for the Solution of Charge Transport in Semiconductors with Applications to Covalent Materials," *Rev. Mod. Phys.*, vol. 55, no. 3, July 1983.

[31] C. W. J. Beenaker, H. van Hoten, "Quantum Transport in Semiconductor Nanostructures," *Solid State Phys.*, vol. 44, no. 1- 228, pp. 1- 111, 1991.

[32] H. Bruus, K. Flensberg, *Many-Body Quantum Theory in Condensed Matter Physics* (Oxford University Press, UK, 2010).

[33] O. Madelung, *Introduction to Solid-State Theory* (Springer-Verlag, Berlin, 1978).

[34] P. Yu, M. Cardona, *Fundamentals of Semiconductors - Physics and Material Properties* (Springer, Berlin, 1999).

[35] A. M. Stoneham, *Theory of Defects in Solids* (Oxford University Press, UK, 1985).

[36] C. R. Helms, E. H. Poindexter, "The Silicon-Silicon-Dioxide System: Its Microstructure and Imperfections," *Rep. Prog. Phys.*, vol. 57, pp. 791 - 852, 1994.

[37] N. Pottier, *Nonequilbrium Statistical Physics* (Oxford University Press, UK, 2010).

[38] J. G. Korvink, A. Greiner, *Semiconductors for Micro- and Nanotechnology* (Wiley-VHC, Morlenbach, Germany, 2002).

[39] T. Grasser, B. Kaczer, *et al.*, "A Two-Stage Model for Negative Bias Temperature Instability," *IEEE 47th Annual International Reliability Physics Symposium*, pp. 33 - 44, April 2009.

[40] T. Grasser, S. Selberherr "Modeling of Negative Bias Temperature Instability," *Journal of Telecommunications and Information Technology*, pp. 92 - 100, February 2007.

[41] S. Bhardwaj, *et al*, "Predictive Modeling of the NBTI Effect for Reliability Design," *IEEE Custom Integrated Circuits Conference*, pp. 189 - 192, September 2006.

[42] T. Grasser, H. Reisinger, *et al*, "Switching Oxide Traps as the Missing Link Between Negative Bias Temperature Instability and Random Telegraph Noise," *IEEE Electron Devices Meeting (IEDM)*, pp. 31.1.1 - 31.1.4, December 2009.

[43] D. M. Fleetwood, *et al*, "Unified Model of Hole Trapping, $1/f$ Noise, and Thermally Stimulated Current in MOS Devices," *IEEE Trans. on Nuclear Science*, vol. 49, no. 6, pp. 2674 - 2683, December 2002.

[44] B. R. Desai, *Quantum Mechanics with Basic Field Theory* (Cambridge University Press, UK, 2010).

[45] R. Shankar, *Principles of Quantum Mechanics* (Springer, New York, 1994).

[46] A. Fetter, J. Walecka, *Quantum Theory of Many-Particle Systems* (Dover, New York, 2003).

[47] J.-P. Colinge, C. Colinge, *Physics of Semiconductor Devices* (Kluwer Academic Press, Boston, 2002).

[48] E. H. Nicollian, J. R. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology* (Wiley, New York, 1982).

[49] N. Weste, K. Eshraghian, *Principles of CMOS VLSI Design - A Systems Perspective* (Addison-Wesley, New York, 1993).

[50] M. Toda, R. Kubo, N. Saitô, *Statistical Physics I - Equilibrium Statistical Mechanics* (Springer-Verlag, Berlin, 1983).

[51] M. Toda, R. Kubo, N. Hashitsume, *Statistical Physics II - Nonequilibrium Statistical Mechanics* (Springer-Verlag, Berlin, 1985).

[52] E. Pop, R. Dutton, K. Goodson, "Thermal Analysis of Ultra-Thin Body Device Scaling," *IEEE Electron Devices Meeting (IEDM)*, pp. 36.6.1 - 36.6.4, December 2003.

[53] J. Lai, A. Majumbar, "Concurrent Thermal and Electrical Modeling of Sub-Micrometer Silicon Devices," *J. Appl. Phys.*, vol. 79, no. 9, pp. 7353 - 7361, May 1996.

[54] S. V. J. Narumanchi, J. Y. Murthy, C. H. Amon, "Boltzmann Transport Equation-Based Thermal Modeling Approaches for Hotspots in Microelectronics," *J. Heat Mass Transfer*, vol. 42, pp. 478 - 491, 2006.

[55] G. K. Wachutka, "Rigorous Thermodynamic Treatment of Heat Generation and Conduction in Semiconductor Device Modeling," *IEEE Trans. on Computer-Aided Design*, vol. 9, no. 11, pp. 1141 - 1149, November 1990.

[56] C. H. Henry, D. V. Lang, "Nonradiative Capture and Recombination by Multiphonon Emission in GaAs and GaP," *Phys. Rev. B*, vol. 15, no. 2, pp. 989 - 1016, January 1977.

[57] M-F. Li, *Modern Semiconductor Quantum Physics* (World Scientific, New York, 1994).

[58] H. Sumi, "Nonradiative Multiphonon Capture of Free Carriers by Deep-Level Defects in Semiconductors: Adiabatic and Non-Adiabatic Limits," *Phys. Rev. B*, vol. 27, no. 4, pp. 2374 - 2386, February 1983.

[59] H. Sumi, "Dynamic Defect Reactions Induced by Multiphonon Nonradiative Recombination of Injected Carriers at Deep Level in Semiconductors," *Phys. Rev. B*, vol. 29, no. 8, pp. 4616 - 4630, April 1984

[60] B. K. Ridley, *Quantum Processes in Semiconductors* (Oxford University Press, UK, 1999).

[61] S. Makram-Ebeid, M. Lannoo, "Quantum Model for Phonon-Assisted Tunnel Ionization of Deep Levels in a Semiconductor," *Phys. Rev. B*, vol. 25, no. 10, pp. 6406 - 6424, May 1982.

[62] S. D. Ganichev, W. Prettl, I. N Yassievich, "Deep Impurity-Center Ionization by Far-Infrared Radiation," Phys. Solid State, vol. 39, no. 11, pp. 1703 - 1726, November 1997.

[63] T. Grasser, *et al*, "Using Six Moments of Boltzmann's Transport Equation for Device Simulation," *J. Appl. Phys.*, vol. 90, no. 5, pp. 2389 - 2396, September 2001.

[64] S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, UK, 1997).

[65] S. Datta, "Nanoscale Device Modeling: the Green's Function Method," *Superlattices and Microstructures*, vol. 28, no. 4, pp. 253 - 278, October 2000.

[66] J.-S. Wang, J. Wang, N. Zeng, "Nonequilibrium Green's Function Approach to Mesoscopic Thermal Transport," *Phys. Rev. B*, vol. 74, 0033408, July 2006.

[67] S. Koswatta, *et al*, "Nonequilibrium Green's Function Treatment of Phonon Scatterng in Carbon-Nanotube Transistors," *IEEE Trans. on Electron Devices*, vol. 54, no. 9, pp. 2339 - 2351, September 2007.

[68] T. Frederiksen, M. Paulsson, M. Brandbyge, A.-P. Jauho, "Inelastic Transport Theory from First Principles: Methodology and Application to Nanoscale Devices," *Phys. Rev. B*, vol. 75, 205413, May 2007.

[69] P. Havu, V. Havu, J. Puska, M. Nieminen, "Nonequilibrium Electron Transport in Two-Dimensional Nanostructures Modeled Using Green's Functions and the Finite-Element Method," *Phys. Rev. B*, vol. 69, 115325, March 2004.

[70] R. Iotti, E. Ciancio, F. Rossi, "Quantum Transport Theory for Semiconductor Nanostructures: A Density-Matrix Formulation," *Phys. Rev. B*, vol. 72, 125347, September 2005.

[71] H. Takeda, N. Mori, "Three-Dimensional Quantum Transport Simulation of Ultra-Small FinFETs," *J. Comput. Electron.*, vol. 4, pp. 31 - 34, 2005.

[72] H. R. Khan, D. Mamaluy, D. Vasileska, "3D NEGF Quantum Transport Simulator for Modeling Ballistic Transport in Nano FinFETs," *Journal of Physics: Conference Series*, IOP Publishing, vol. 107, 2008.

[73] H. R. Khan, D. Mamaluy, D. Vasileska, "Fully 3D Self-Consistent Quantum Transport Simulation of Double-gate and Tri-gate 10 nm FinFETs," *J. Comput. Electron.*, vol. 7, pp. 346 - 349, February 2008.

[74] H. R. Khan, D. Mamaluy, D. Vasileska, "Quantum Transport Simulation of Experimentally Fabricated Nano-FinFET," *IEEE Trans. Electron Devices*, vol. 54, no. 4, pp. 784 - 796, April 2007.

[75] H. R. Khan, D. Mamaluy, D. Vasileska, "Approaching Optimal Device Characteristics of 10 nm High-Performance Devices: A Quantum Transport Simulation Study of Si FinFET," *IEEE Trans. Electron Devices*, vol. 55, no. 3, pp. 743 - 753, March 2008.

[76] X. Shao, Z. Yu, "Nanoscale FinFET Simulation: A Quasi-3D Quantum Mechanical Model Using NEGF," *Solid-State Electronics*, vol. 49, pp. 1435 - 1445, June 2005.

[77] W. H. Press, *et al*, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, UK, 2007).

[78] L. Villablanca, "Mesh Generation Algorithms for Three-Dimensional Semiconductor Process Simulation," *Series in Microelectronics*, vol. 97, 2000.