Relaxation Effects in MOS Devices due to Tunnel Exchange with Near-Interface Oxide Traps

by

Theodore L. Tewksbury III

S.B. Massachusetts Institute of Technology (1983)
S.M. Massachusetts Institute of Technology (1987)

Submitted in Partial Fulfillment
of the Requirements for the
Degree of

Doctor of Philosophy

in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

June 1992

© Massachusetts Institute of Technology 1992
All rights reserved

Signature of Author _____
Department of Electrical Engineering and Computer Science
May 26, 1992

Certified by _____
Hae-Seung Lee
Associate Professor of Electrical Engineering
Thesis Supervisor

Accepted by _____
Campbell L. Searle, Chairman
Committee on Graduate Students

ARCHIVES
MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

JUL 10 1992

LIBRARIES

Relaxation Effects in MOS Devices due to Tunnel Exchange with Near-Interface Oxide Traps

by

Theodore L. Tewksbury III

Submitted to the
Department of Electrical Engineering and Computer Science
on May 26, 1992 in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy.

Abstract

Transistors and capacitors in many types of analog MOS circuits such as charge-redistribution A/D converters, switched-capacitor filters and sample-and-holds, are routinely subjected to momentary large-signal stress conditions during the course of their normal operation. During these transients, charge can be captured by oxide traps close to the Si-SiO_2 interface by tunneling. When the stress is removed, the emission of this trapped charge persists for periods which greatly exceed the capture time, leading to dielectric relaxation in capacitors and transient threshold voltage shifts in MOSFETs. While the amount of trapped charge is too small to be of interest in most digital circuits where the resulting threshold voltage shifts fall comfortably within noise margins, in analog circuits it can lead to hysteresis effects, charge redistribution errors and long-tail components in the settling time of comparators and operational amplifiers with MOS input stages. As analog integrated circuits are pushed towards higher levels of speed and accuracy, these trapping-induced relaxation effects will impose increasingly important performance limitations.

Previous experimental studies of charge trapping by near-interface oxide traps have neglected the voltage and time ranges relevant to analog circuits, while existing theories do not provide a complete description of the various mechanisms by which traps can be charged and discharged. This thesis undertakes a detailed characterization of trapping-induced relaxation effects in MOSFETs and capacitors as a function of bias, time, temperature and geometry, under conditions relevant to analog circuit operation. A theoretical description of various relaxation mechanisms is presented and it is shown that the dominant charging/discharging mechanism under the bias conditions of interest is by the elastic tunneling of charge carriers to and from oxide traps. Tunneling occurs directly between the silicon band edges and an oxide trap, or in conjuction with either lattice relaxation multiphonon emission in the insulator or the assistance of fast surface states at the interface. Experimental data from transistors with conventional and nitrided dielectrics are presented and explained using

this model. The relation of threshold voltage relaxation to 1/f noise and the effect of correlated mobility fluctuations are also discussed.

Techniques for the modeling and minimization of relaxation effects in circuits are presented.

Thesis Supervisor: Hae-Seung Lee
Title: Associate Professor of Electrical Engineering

Acknowledgments

As in any scientific endeavor, this work would not have been possible without the generous assistance of a great number of people. Most importantly, I must thank my thesis advisor, Prof. Hae-Seung Lee, for the technical, managerial and personal support that he has provided over the years. His contributions to this research were invaluable and the influence he had on my thinking will long outlast it. I would also like to thank Harry for his genuine concern for my future and for the professional advice he has offered over the years, which have played a pivotal role in shaping my career.

I am also tremendously grateful to my thesis readers: Prof. Stephen Senturia for his crucial contributions to the physics and for instilling in me a healthy objectivity and appreciation for scientific rigor. His inspiration and example are responsible for my being in this field in the first place. Much thanks to Prof. Charles Sodini, whose contagious enthusiasm and sense of humor have always served as reminders in times of discouragement that we are, after all, in this for the love of it. His incisive questions and incomparable insight into MOS device physics have always kept the research on track and relevant.

The generosity of Brad Scharf, Steve Lewis and P. J. Garrivan at Analog Devices Semiconductor, who supplied wafers for the threshold voltage measurements, is greatly appreciated. Thanks also to John Fattaruso and Richard Hester at Texas Instruments for providing test chips for the capacitor characterization and to Jeff Gross at MTL for donating samples from his ROXNOX process.

This research would never have been begun were it not for the efforts of a number of people at Analog Devices who made it possible for me to return to MIT under their continued support. My sincere appreciation goes to D. G. Ganesan, John Fernandes and Peter Real who, by creative management, helped to make this possible. I am also grateful to Gerry Miller and Mike Timko for sharing their own experience with relaxation effects in integrated circuits, to Peter Casparius for invaluable technical assistance during the measurements and to Tom Lee, Francisco Dos Santos and others too innumerable to mention who listened patiently to my ravings and kept me sane.

Thanks to my friends and colleages — Gee Rittenhouse and Kathy Krisch for many useful discussions on device physics and to my officemates Mikko Hakkarainen and John Bulzacchelli for listening to my various theories and offering helpful criticism. I am indebted to Joe Kung, Curtis Tsai, Vincent Loh and Patrice Parris for their help on various computer-related problems which periodically threatened to bring my work to a halt. Thanks to all the other students and staff at MTL, too innumerable to mention, who make it a stimulating and enjoyable place to work.

Contents

1 Introduction 1
- 1.1 Description of the Problem . . . 1
  - 1.1.1 Threshold Voltage Relaxation in MOSFETs . . . 2
  - 1.1.2 Dielectric Relaxation in Capacitors . . . 4
- 1.2 Organization of the Thesis . . . 5

2 Background 9
- 2.1 Polarization Mechanisms . . . 9
  - 2.1.1 Mobile Ions . . . 10
  - 2.1.2 Dipolar/Interfacial Polarizations . . . 11
  - 2.1.3 Fast Surface States . . . 12
  - 2.1.4 Oxide Traps . . . 13
  - 2.1.5 Defect Relaxation . . . 14
  - 2.1.6 Hot Carrier Effects . . . 14
  - 2.1.7 Thermal Hysteresis . . . 15
- 2.2 Previous Studies . . . 16
  - 2.2.1 Small-Signal Measurements . . . 16
  - 2.2.2 Large-Signal Measurements . . . 17
- 2.3 The Model of Heiman and Warfield . . . 20

3 Theory of Tunneling to and from Oxide Traps 23
- 3.1 Introduction . . . 23
- 3.2 Fundamental Mechanisms . . . 24
  - 3.2.1 Insulator Band-to-Trap Transitions . . . 24
  - 3.2.2 Tunneling to and from Oxide Traps . . . 26
- 3.3 Tunneling Time Constant . . . 31
- 3.4 Tunneling between Oxide Traps and Band States . . . 34
  - 3.4.1 Elastic Tunneling . . . 35
  - 3.4.2 Inelastic Tunneling . . . 37
  - 3.4.3 Lattice Relaxation Multiphonon Emission (LRME) . . . 38
- 3.5 Tunneling between Oxide Traps and Interface States . . . 42

3.5.1 Tunneling Time Constant 42
3.5.2 Two-step Process 44
3.5.3 Assumptions 47
3.5.4 Numerical Solution 49
3.5.5 Approximate Solutions 50
3.6 Two-Band Model 58
3.6.1 Attenuation Constant in the Oxide 59
3.6.2 Attenuation Constant in the Semiconductor 62
3.7 Summary 62
3.8 Discussion 64

4 Threshold Voltage Relaxation in MOSFETs 67
4.1 Introduction 67
4.2 Formulation of the Problem 68
4.2.1 Threshold Voltage Shift 69
4.2.2 Correlated Mobility Fluctuations . . . . . 6[illegible]
4.2.3 Equivalent Gate-Source Voltage Shift 70
4.2.4 Effective Threshold Voltage Shift 71
4.3 General Solution 72
4.3.1 The Effect of Band Bending in the Oxide 72
4.3.2 Trapping Components 74
4.4 Approximate Voltage and Time Dependence 87
4.4.1 Two-Step Process 89
4.4.2 Lattice Relaxation 101
4.5 Temperature Dependence 103
4.5.1 Correlated Mobility Fluctuations 103
4.5.2 Carrier Distributions in the Semiconductor 104
4.5.3 Two-Step Process 107
4.5.4 Lattice Relaxation 108
4.5.5 Summary 109
4.6 Geometry Dependence 109
4.7 Hysteresis Effects 111
4.7.1 Charging and Emission Kinetics 113
4.7.2 Time Constant Asymmetries 114
4.8 Dynamic Effects 116
4.9 Summary 117

5 Dielectric Relaxation in Capacitors 125
5.1 Introduction 125
5.2 Empirical Models 127
5.3 Mechanisms 128
5.4 Trapping Theory 131

5.4.1 Trapping from the Bottom Plate . . . 132
5.4.2 Trapping from the Top Plate . . . 137
5.5 Model . . . 139
5.5.1 Simple Analysis . . . 142
5.5.2 Exact Analysis . . . 144
5.6 Summary . . . 145

6 Measurement Techniques **147**
6.1 Introduction . . . 147
6.2 Threshold Voltage Relaxation . . . 148
6.2.1 Introduction . . . 148
6.2.2 Measurement Circuit . . . 150
6.2.3 Conditions for Constant Power Dissipation . . . 154
6.3 Dielectric Relaxation in Capacitors . . . 156

7 Experimental and Simulation Results **159**
7.1 Introduction . . . 159
7.2 Threshold Voltage Relaxation . . . 159
7.2.1 Thermal Considerations . . . 161
7.2.2 Dependence on Device Characteristics and Test Conditions . . 163
7.2.3 Nitrided and Reoxidized Nitrided Gate Oxides . . . 191
7.2.4 Simulation Results and Extraction of Trap Densities . . . 202
7.2.5 Relation to 1/f Noise . . . 218
7.2.6 Circuit Issues . . . 223
7.3 Dielectric Relaxation in Capacitors . . . 231

8 Conclusions **237**
8.1 Summary . . . 237
8.1.1 Theory . . . 237
8.1.2 Experiment . . . 239
8.1.3 Discussion . . . 239
8.2 Suggestions for Future Work . . . 244

A Tunneling Transition Probability **249**
A.1 Introduction . . . 249
A.2 The Transfer Hamiltonian Formalism . . . 250

B Tunneling Time Constants **255**
B.1 Band-to-Trap Tunneling . . . 256
B.1.1 One-dimensional Oxide Trap Potential . . . 258
B.1.2 Three-dimensional Oxide Trap Potential . . . 263
B.2 Interface State to Oxide Trap Tunneling . . . 268

B.2.1 One–Dimensional Oxide Trap Potential 273
B.2.2 Three Dimensional Oxide Trap Potential 275

C Hole Tunneling and Lattice Relaxation **277**

D Solution of Coupled Rate Equations **281**

E Band bending Calculations **283**

F Recovery Voltage for Donor-like Traps **287**

G List of Symbols **291**

H Abbreviations and Acronyms **297**

List of Figures

1.1 Observed drain current undershoot and relaxation in an NMOS transistor stressed with a postive gate voltage pulse. 2

1.2 Trapping-induced threshold voltage shift (a) Before stress; oxide traps empty (b) During stress; traps fill with electrons (c) Relaxation; emission of electrons from traps. 3

1.3 Dielectric relaxation in a capacitor due to a distribution of traps close to the bottom plate (a) Charged to voltage V_p for time t_s, during which electrons are trapped from the bottom plate (b) Discharged to ground for time t_d, insufficient for all traps to discharge (c) Plates are open-circuited and a recovery voltage $V_r(t)$ develops as trapped charge is emitted to the bottom plate. 5

2.1 (a) Dipolar polarization (b) mobile ions (c) trapping by interface states or near-interface oxide traps 11

3.1 Mechanisms for introducing charge carriers into SiO_2. (a) Thermal capture of electrons injected into the insulator conduction band; (1) avalanche electron injection (AEI), (2) Fowler-Nordheim tunneling injection (FNTEI), (b) Direct tunneling to and from oxide traps; (3) elastic tunneling from the conduction band, (4) inelastic (phonon-assisted) tunneling, (5) elastic tunneling followed by cascade phonon emission or lattice relaxation multiphonon emission, (6) two-step, interface state-assisted tunneling. 27

3.2 Elastic tunneling followed by cascade phonon emission. 29

3.3 Configuration coordinate diagram showing (a) system consisting of an electron in the conduction band and an empty oxide trap (b) system consisting of an electron in a bound state of an oxide trap. 30

3.4 Band diagram showing transitions between states at energy E_l at the interface and states at energy E_r in the oxide 32

3.5 Elastic tunneling from the conduction band. (a) Before stress; traps empty, (b) During stress; traps fill by tunneling from filled conduction band states, (c) After stress; electrons tunnel out to empty conduction band states. 35

3.6 Interface trap-assisted tunneling. (a) Before stress, (b) During stress; oxide traps fill by tunneling from interface states, (c) After stress; oxide traps empty by tunneling to empty interface states. 43

3.7 Two-step tunneling between oxide traps and interface states showing possible electron transitions. 44

3.8 Simulations showing oxide trap fractional occupancy as a function of time for $V_G(initial) = 5\text{V}, V_G(final) = 1V$, $D_{it} = 1.0\times10^{10}\text{cm}^{-2}\ \text{eV}^{-1}$, $D_{ot} = 1.0 \times 10^{17}\text{cm}^{-3}\ \text{eV}^{-1}$, $x = 10\mathring{A}$ and (a) $E_t = -0.4\text{eV}$, $T = 25°\text{C}$ (b) $E_t = -0.4\text{eV}$, $T = -55°\text{C}$ (c) $E_t = -0.1\text{eV}$, $T = -55°\text{C}$. 55

3.9 Simulations showing oxide trap fractional occupancy as a function of time for $V_G(initial) = 5\text{V}, V_G(final) = 1V$, $D_{it} = 1.0\times10^{10}\text{cm}^{-2}\ \text{eV}^{-1}$, $D_{ot} = 1.0\times10^{17}\text{cm}^{-3}\ \text{eV}^{-1}$, $E_t = -0.4\text{eV}, T = -55°\text{C}$ and (a) $x = 15\mathring{A}$ (b) $x = 10\mathring{A}$ (c) $x = 5\mathring{A}$. 56

3.10 Simulations showing oxide trap fractional occupancy as a function of time for $V_G(initial) = 5\text{V}, V_G(final) = 1V$, $D_{it} = 1.0\times10^{10}\text{cm}^{-2}\ \text{eV}^{-1}$, $E_t = -0.4\text{eV}$, $x = 5\mathring{A}$, $T = -55°\text{C}$ and (a) $D_{ot} = 1.0 \times 10^{17}\text{cm}^{-3}\ \text{eV}^{-1}$ (b) $D_{ot} = 1.0 \times 10^{18}\text{cm}^{-3}\ \text{eV}^{-1}$ (c) $D_{ot} = 1.0 \times 10^{19}\text{cm}^{-3}\ \text{eV}^{-1}$. 57

3.11 Simple two band model for the insulator. $k^2 < 0$ in the energy gap corresponding to decaying exponentials. 60

4.1 Energy band diagram of the MOS structure showing definition of coordinates for measurement of oxide trap energies; (a) flatband (b) with applied electric field $\mathcal{E}$. 73

4.2 The nine possible combinations of direct and interface state assisted tunneling transitions by which an oxide trap can communicate with the silicon bands during a pulsed gate voltage measurement. 76

4.3 Examples of additional tunneling channels made possible by lattice relaxation multiphonon emission. 77

4.4 Band diagrams for an NMOS transistor subjected to positive gate voltage pulses, $V_p > 0$. (a) During stress, showing mechanism by which traps charge (b) After stress, showing mechanism by which traps empty, and classification of trap energy levels by fill mode/empty mode. 80

4.5 Classification of oxide trap energy levels into mutually exclusive regions in (E_t, x) space according to fill mode/empty mode for positive gate voltage pulses $V_p > 0$. 82

4.6 Band diagrams for an NMOS transistor subjected to negative gate voltage pulses, $V_p < 0$. (a) During stress, showing mechanisms by which traps charge (b) After stress, showing classification of trap energy levels by fill mode/empty mode. 85

4.7 Classification of oxide trap energy levels into mutually exclusive regions in (E_t, x) space according to fill mode/empty mode for negative gate voltage pulses $V_p < 0$. 86
4.8 Band diagram for an NMOS transistor pulsed with a positive gate voltage. 89
4.9 Spatial variation of the filled oxide trap density with time (a) $g(x, t_s)$ during stress, (b) $g(x, t_s)[1 - g^o(x, t_r)]$ during emission, after $t_s = 10$ms. 91
4.10 Simulated relaxation transient showing components *it/it* (Eq. 4.68), *it/cb* (Eq. 4.77) and *cb/cb* (Eq. 4.81). 94
4.11 Relaxation transient of Fig. 4.10 plotted on (a) log-linear axes (b) log-log axes. 95
4.12 Exact and Boltzmann approximations for the surface potential versus gate voltage for a 250$\AA$ gate oxide with substrate doping of $N_A = 1.0 \times 10^{15}\text{cm}^{-3}$. 96
4.13 Components of threshold voltage shift versus stress time t_s. 97
4.14 Components of threshold voltage shift versus stress voltage V_p. 98
4.15 Components of threshold voltage shift versus stress voltage V_p using lattice relaxation multiphonon emission model. 102
4.16 (a) The functions f and f_0 and the difference $\Delta f = f - f_0$ for a trap at $x = 0$ (b) The function $\Delta f(E) = f(E) - f_0(E)$ as a function of temperature. 106
4.17 The function $\Delta f(E) = f(E) - f_0(E)$ for negative stress voltages. . . 107
4.18 Plan view of a MOSFET defining terms used in derivation of geometry dependence 110
4.19 Fractional occupation of oxide traps $g(x, t_s)[1 - g(x, t_r)]$ at various depths as a function of time during stress and emission 113
4.20 Threshold voltage shift as a function of stress time and relaxation time for $t_{ox} = 250\AA$ and $D_{ot} = 1 \times 10^{17}$ cm^{-3}. 114
4.21 Charging and Emission transients obtained by integrating Eq. (a) Solid line: $\tau_f = \tau_e$ (b) Dashed line: $\tau_e = 10\tau_f$ (c) Dotted line: $E_{Be} = E_{Bc} + 1eV$ 115
4.22 Threshold voltage shift as a function of time for 1MHz pulse rate . . 117

5.1 Dow's model for dielectric absorption in a capacitor. 128
5.2 Energy band diagrams corresponding to dielectric relaxation measurement sequence: (a) initial (b) charge (c) discharge (d) float. 132
5.3 (i) Acceptor traps filled by tunneling from the bottom plate during positive stress, and plots of (ii) charge density (iii) electric field (iv) potential in the oxide. (a) Immediately following discharge period [$t_r = 0^+$] (b) Float period [$t_r > 0$]. 133

5.4 (i) Acceptor traps emptied by tunneling to the bottom plate during negative stress, and plots of (ii) charge density (iii) electric field (iv) potential in the oxide. (a) Immediately following discharge period [$t_r = 0^+$] (b) Float period [$t_r > 0$]. 136
5.5 (i) Acceptor traps emptied by tunneling to the top plate during positive stress, and plots of (ii) charge density (iii) electric field (iv) potential in the oxide. (a) Immediately following discharge period [$t_r = 0^+$] (b) Float period [$t_r > 0$]. 138
5.6 (i) Acceptor traps filled by tunneling from the top plate plate during negative stress, and plots of (ii) charge density (iii) electric field (iv) potential in the oxide. (a) Immediately following discharge period [$t_r = 0^+$] (b) Float period [$t_r > 0$]. 140
5.7 Spatial distribution of trapped charge during the dielectric absorption measurement sequence. (a) $t = t_s$ (b) $t = t_s + t_d$ (c) $t = t_s + t_d + t_r$ (d) traps which contribute to the relaxation integral for $t_r > 0$. 142

6.1 Mechanisms which may contribute to threshold voltage shifts and experimental conditions by which they are elimated. 151
6.2 Open-loop measurement circuit. 152
6.3 Closed-loop measurement circuit. 153
6.4 I-V characteristics before and after stress and constant power dissipation characteristic 155
6.5 Dielectric relaxation measurement circuit; (a) simplified circuit schematic (b) timing diagram. 157

7.1 Oscilloscope photographs showing measured V_T shifts in process 5 ($W/L = 20\mu m/5\mu m$); $t_s = 10ms$, $V_p = \pm 5V$, V_{DS}=350mV. Vertical scale = 100 μV per division (input-referred), Horizontal scale = 20ms per division. 160
7.2 Drain current vs. temperature (process 2) 163
7.3 (a) Dependence of V_T shift on drain voltage, (b) change in power dissipation during stress as a function of drain voltage. 164
7.4 Drain voltage dependence of initial V_T shift in process 3. 165
7.5 Relaxation transients for process 2 (a) comparison of NMOS and PMOS (b) close-up view of the PMOS V_T shift in (a). 167
7.6 NMOS and PMOS relaxation transients for (a) process 3 (b) process 5. 168
7.7 Threshold voltage relaxation transients plotted on a log time scale with stress voltage as a parameter; (a) $t_s = 10\mu s$ (b) $t_s = 10ms$. 169
7.8 Relaxation measured with open loop circuit. 170
7.9 Dependence of threshold voltage relaxation on stress time. 171
7.10 Initial threshold voltage shift vs. stress time for two different processses. 171
7.11 Initial amplitude of threshold voltage shift versus stress time. 172

7.12 Initial V_T shift versus stress voltage for Process 1. 173
7.13 (a) Stress voltage dependence (b) Surface potential versus gate voltage (c) Change in power dissipation during stress [Process 2]. 174
7.14 Energy band diagrams for positive (top) and negative (bottom) stress sequences on an NMOS transistor, showing field emission of holes trapped during negative stress. 176
7.15 Threshold voltage relaxation transients with initial gate-source voltage as a parameter. 178
7.16 Initial threshold voltage shift vs. stress voltage with initial gate-source voltage as a parameter. 179
7.17 Initial threshold voltage shift vs. stress voltage with initial gate bias as a parameter [Process 1] (a) $t_s = 10$ms (b) $t_s = 10\mu$s. 180
7.18 Threshold voltage relaxation transients at $V_{BS} = 0$V and $V_{BS} = -5$V. 181
7.19 Initial threshold voltage shift vs stress voltage with substrate bias as a parameter. 182
7.20 Band diagram for an NMOS transistor with $V_{SB} > 0$ (a) pre-stress $V_p = 0$ (inversion) (b) small negative stress V_p (depletion) (c) large negative stress V_p (accumulation). 183
7.21 Threshold voltage relaxation transients at several different temperatures. 185
7.22 Initial amplitude of threshold voltage shift versus temperature at (a) constant current (b) constant drain-source voltage. 186
7.23 Initial threshold voltage shift versus temperature for several different processes (NMOS 20/5); $V_{GS} = 1.2V$, $V_{DS} = 500$mV, $t_s = 10$ms. . . . 187
7.24 Initial threshold voltage shift versus stress voltage with temperature as a parameter. 187
7.25 Initial amplitude of threshold voltage shift versus stress voltage at $V_{BS} = 0$V and $V_{BS} = -3$V at (a) -30°C and (b) 125°C. 188
7.26 Channel length dependence 189
7.27 Channel length dependence (Process 3) 190
7.28 Channel width dependence 190
7.29 Oscilloscope photographs showing measured V_T shifts in process 7 ($W/L = 20\mu$m/5μm); V_{DS}=350mV, $t_s = 10$ms, $V_p = \pm 2$V. Vertical scale = 500μV per division (input-referred), Horizontal scale = 20ms per division. 192
7.30 Threshold voltage relaxation transients for 100Å oxides, nitrided oxides and reoxidized nitrided oxides; (a) NMOS (b) PMOS. 193
7.31 Threshold voltage relaxation transients for nitrided oxides (a) NMOS (b) PMOS 195
7.32 Initial threshold voltage shift vs. stress time with stress voltage as a parameter, (NOX 20/5) (a) NMOS (b) PMOS. 196

7.33 Initial threshold voltage shift versus stress voltage with stress time as a parameter (NMOS); (a) control oxide (b) NOX (c) ROXNOX. . . . 197
7.34 Initial threshold voltage shift versus stress voltage with stress time as a parameter (PMOS); (a) control oxide (b) NOX (c) ROXNOX. . . . 198
7.35 Initial threshold voltage shift vs. stress voltage with drain-source voltage as a parameter; NOX NMOS 20/5, $V_{GS} - V_T = 700$mV. 199
7.36 Initial threshold voltage shift vs. stress voltage with initial gate bias as a parameter, NOX NMOS 20/5, V_{DS}=200mV (a) stress time = 10ms (b) stress time = $1\mu s$. 200
7.37 Initial threshold voltage shift vs. stress voltage with source-substrate voltage as a parameter. 201
7.38 Threshold voltage relaxation transients at several different temperatures.201
7.39 $\Delta V_T(1ms)$ vs. stress voltage; (a) Simulation using two-step model (b) Oxide trap distribution used in (a). 203
7.40 $\Delta V_T(t_d = 1ms)$ vs. stress voltage; (a) Simulation using lattice relaxation multiphonon emission (LRME) model (b) Oxide trap distribution used in (a). 204
7.41 (a) Relaxation transients at three temperatures (b) Decomposition by component. 210
7.42 Temperature dependence in strong inversion (a) Two-step model, (b) Lattice relaxation MPE model. 211
7.43 Simulated and measured $\Delta V_T(t_d = 20\mu s)$ vs. V_p characteristics at two different temperatures. 212
7.44 Comparison of models to temperature dependence of ΔV_T vs. V_p; (a) T= 125°C (b) T= −30°C. 213
7.45 (a) Initial gate bias dependence; simulated and measured. (b) Change in trap occupancies as a function of gate bias. 215
7.46 Simulated and measured V_T shifts versus stress time with stress voltage V_p as a parameter. 216
7.47 Simulated relaxation transient showing decomposition by mechanism. 216
7.48 Dependence of initial threshold voltage shift on pulse amplitude, NOX NMOS 20/5, $V_{GS} - V_T = 1.1V$, $V_{DS} = 200$mV (a) Breakdown by component (b) Parameterized by stress time. 217
7.49 Comparison of three different 20/5 NMOS transistors (a) threshold voltage relaxation transient (b) 1/f noise. 220
7.50 1/f noise at 100 Hz and peak V_T shifts for various processes (a) NMOS (b) PMOS. 222
7.51 Simple SPICE subcircuit model for V_T relaxation. 224
7.52 Measured V_T relaxation data and simulations using SPICE model (t_s = 10ms). 225

7.53 Simulated pulse response of a differential pair using the SPICE V_T relaxation model. 226
7.54 Measured and simulated relaxation transients in process 2 using the SPICE model ($t_s = 10\mu$s). 227
7.55 Accuracy (in bits) versus clock frequency with reference voltage as a parameter. 228
7.56 Suppression of V_T shift by operating in strong inversion with a backgate bias (a) NMOS (b) PMOS. 230
7.57 Recovery voltage vs. relaxation time for a capacitor with a LPCVD SiO_2. 231
7.58 Recovery voltage vs. relaxation time for a capacitor with a nitrided oxide dielectric. 232
7.59 Recovery voltage vs. relaxation time with stress time (t_s) as a parameter. 232
7.60 Recovery voltage vs. relaxation time with discharge time (t_d) as a parameter. 233
7.61 Temperature dependence of the recovery voltage. (a) relaxation transients parameterized by temperature (b) asymptotic recovery voltage vs. temperature. 234

B.1 Energy band diagram showing an electron tunneling from a conduction band state, on the left side of the insulator potential barrier, into an oxide trap on the right. 256
B.2 Left and right side systems for use in Bardeen's method 257
B.3 Time constant prefactor (a) Electric field dependence (b) Energy dependence. 262
B.4 Time constant versus energy above the conduction band for a trap at a depth of 50$\AA$ from the interface, using Eq. B.69 (solid line), Eq. B.71 (dotted line) and Eq. B.74 (dashed line) 268
B.5 Double-well potential 269
B.6 Band diagram showing interface and oxide trap potential wells. 272
B.7 Separation of the exact Hamiltonian into two partial Hamiltonians for (a) left side wavefunction (interface state) (b) right side wavefunction (oxide trap) 273

List of Tables

4.1 Trapping components for positive stress and $x < x_1$ 81
4.2 Trapping components for positive stress and $x_1 < x < x_2$ 83
4.3 Trapping components for negative stress and $x < x_1$ 84
4.4 Trapping components for negative stress and $x_1 < x < x_2$ 86
4.5 Summary of First-Order Time Constants 120
4.6 Limiting Forms of Time Constants $\tau_{ot/j}(E_t, x)$ 123

7.1 Process Summary 161
7.2 NMOS ($t_s = 10$ms, $t_d = 1$ms) 162
7.3 PMOS ($t_s = 10$ms, $t_d = 1$ms) 162
7.4 NMOS ($t_s = 10$ms, $t_d = 1$ms, $V_{GS} - V_T = 500$mV, $V_{DS} = 300$mV, W/L=20/5) 191
7.5 PMOS ($t_s = 10$ms, $t_d = 1$ms, $V_{GS}-V_T$=-700mV, V_{DS}=300mV, W/L=20/5)191
7.6 Model Parameters used in Simulations 208
7.7 Comparison of measured V_T shifts with estimates based on 1/f noise. 223

Chapter 1

Introduction

1.1 Description of the Problem

The performance of MOS devices in circuits depends critically on the stability of device characteristics with time. Traps in the insulator and at the semiconductor-insulator interface are a well known source of such instabilities, causing threshold voltage shifts and degradation of transconductance and subthreshold characteristics. These traps can be caused by dangling bonds and defects created during device fabrication or under high electric fields due to the breaking of hydrogen bonds or strained intrinsic electron-pair bonds by energetic electrons and holes [1].

Previous studies of the effects of oxide traps on MOS device characteristics can be classified according to the time or frequency scales of interest, the mechanisms by which the traps are charged and discharged, and the excitation signals – large-signal or small-signal – under which the devices are tested. Long-term instabilities, on the order of hours to years, are of interest for the reliability and lifetime of MOS devices and have been heavily researched, generally under accelerated conditions at voltages greatly exceeding those encountered in normal circuit operation. Large threshold voltage drifts in III-V and II-VI MISFETs have been observed at normal operating voltages on time scales of 1-10 seconds. However, short-term changes in MOS device characteristics on the microsecond to millisecond time scales relevant to the performance of devices in circuits have until now received very little attention.

The mechanisms by which electrons are captured and emitted from oxide traps depend strongly on electric field. Oxide traps can be charged and discharged either by the thermal capture of carriers which have been injected into the oxide bands or by direct quantum mechanical tunneling. The former mechanism, which requires high electric fields, has been researched extensively in relation to MOS degradation and reliability and as a tool for the study of insulating films. Carriers may be injected into the oxide bands by a variety of mechanisms, including avalanche injection (AEI) and Fowler-Nordheim tunneling (FN-TEI) under high vertical electric fields or as hot

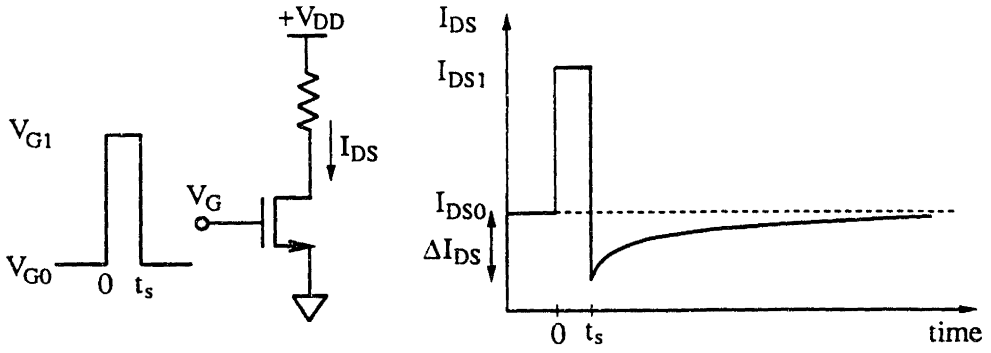


Figure 1.1: Observed drain current undershoot and relaxation in an NMOS transistor stressed with a postive gate voltage pulse.

electrons created by impact ionization under the high transverse electric fields at the drain end of the channel in a saturated MOSFET. At electric fields which are too low for the injection of carriers into the oxide bands, the charging and discharging of oxide traps can occur by the direct tunneling of electrons or holes through the interface potential barrier.

Small-signal manifestations of charge exchange with traps in the oxide and at the interface are well known. The $1/f$ noise in MOS devices is believed to be due to fluctuations in carrier number and mobility resulting from the direct tunneling of charge to and from oxide traps [2] in response to small-signal gate voltage excursions. However, the corresponding problem in the large-signal regime, at voltages comparable to those encountered in analog circuits, but not so large as to cause the injection of carriers over the barrier, has been largely overlooked.

This thesis investigates the effects of oxide traps under large-signal stress conditions at times and voltage ranges relevant to analog circuit operation. Voltages in this regime are too low for the generation of oxide traps or for the injection of charge carriers into the SiO_2 so that charge exchange can only occur by the direct tunneling of electrons and holes to and from pre-existing oxide traps. The relaxation effects which occur due to charge trapping under large-signal stress conditions have an immediate impact on the short-term performance of MOS devices and impose fundamental limitations on the performance of future high-speed and high-accuracy analog circuits.

1.1.1 Threshold Voltage Relaxation in MOSFETs

Consider an NMOS transistor subjected to a momentary positive gate voltage pulse, as shown in Fig. 1.1. When the pulse is removed, the drain current first undershoots and then slowly relaxes back to its initial value. In response to a negative gate voltage pulse, there is an overshoot of the drain current with respect to its initial

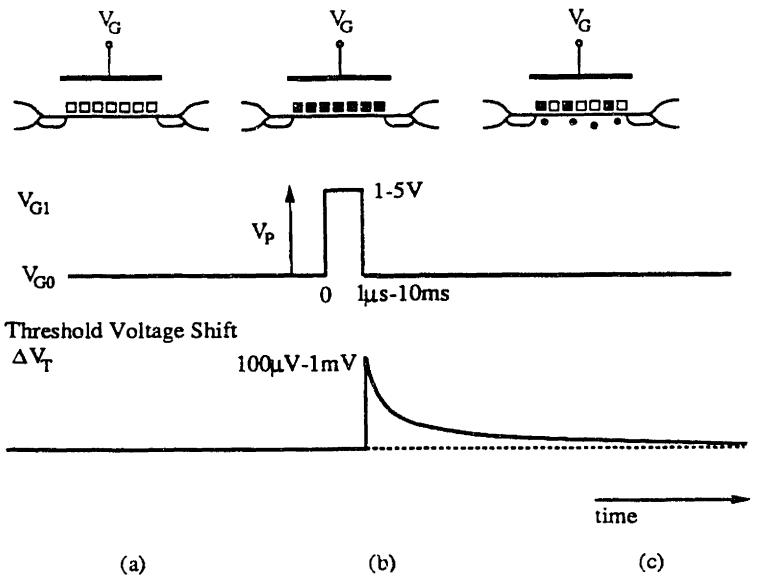


Figure 1.2: Trapping-induced threshold voltage shift (a) Before stress; oxide traps empty (b) During stress; traps fill with electrons (c) Relaxation; emission of electrons from traps.

value, although the amplitude is generally smaller than for positive pulses. A similar effect occurs in PMOS devices, but the amplitude of the overshoot is much less, generally by about an order of magnitude, than that observed in NMOS transistors.

The change in drain current in response to a gate voltage pulse ΔV_G is equivalent to an effective threshold voltage shift, of polarity such that $\Delta V_T/\Delta V_G > 0$. This effect occurs under normal circuit bias conditions, over a range of drain-source voltages, and in the linear as well as in the saturation region. The threshold voltage shift decays with a very long tail, over time scales much longer than those required to produce it.

Various mechanisms can account for this type of behavior and these will be discussed in this thesis. Most of these mechanisms, however, can be eliminated or greatly minimized by appropriate bias conditions and circuit design. The one mechanism which cannot be completely eliminated, and which comprises the topic of this thesis, is the tunneling of charge to and from near-interface oxide traps. During transient stress conditions, oxide traps can capture inversion layer charge carriers by tunneling, as shown in Fig. 1.2. When the stress is removed, this charge is reemitted from the traps with a broad distribution of time constants, causing a shift in the threshold voltage which relaxes over a very long period of time.

In typical MOS analog circuits such as switched-capacitor filters [3, 4, 5, 6] and charge-redistribution A/D converters [7, 8, 9], the differential inputs of operational amplifiers or comparators are routinely subjected to momentary asymmetrical voltage stresses during their normal operation. In comparators, the magnitude of the

stress can be as large as half the full scale voltage of the A/D converter, while in operational amplifiers it can be on the order of the supply voltage under slewing conditions. Threshold voltage shifts which occur under these transient stress conditions due to tunnel exchange with oxide traps can cause offset voltage shifts and long-tail components in the settling time of MOS comparators and opamps, This, in turn, can lead to linearity errors in MOS A/D and D/A converters. The gradual emission of charge from oxide traps can also cause a slow discharge component from a MOS switch in the off-state which may be collected by nearby p-n junctions, causing errors at high-impedance circuit nodes.

The amplitude of trapping-induced threshold voltage shifts are generally on the order of 100μV to 1 mV, too small to be of importance in most digital circuits where they fall well within noise margins. However, in analog circuits which are sensitive to the actual value of the threshold voltage at any given time, the time-dependent threshold voltage shifts can have an important impact on circuit performance, particularly at the 14 bit level and beyond. In the past, the problem of threshold voltage relaxation has received very limited attention from the circuit design community [10, 11, 7], probably because it is commonly masked by other larger sources of error in circuits, such as charge injection [12, 13, 14], noise and thermal drift. However, as A/D and D/A converters are pushed towards higher levels of performance and these other sources of error are eliminated by self calibration [7, 15], correlated double sampling [16] and other circuit techniques, threshold voltage relaxation will impose an increasingly important limitation on the accuracy and speed that can be obtained from analog integrated circuits.

1.1.2 Dielectric Relaxation in Capacitors

The tunneling of charge to and from oxide traps can also cause dielectric relaxation in capacitors. Consider a capacitor which is charged to a constant voltage for some period of time and then momentarily discharged to ground. When the capacitor plates are subsequently open-circuited, the voltage across an ideal capacitor will remain zero. In real capacitors, however, the voltage often exhibits a tendency to drift back in the direction of the original charging voltage as shown in Fig. 1.3. This *recovery voltage* is the result of dielectric relaxation processes within the insulator. This thesis will examine one possible relaxation mechanism – the tunneling of charge to and from oxide traps, as shown schematically in Fig. 1.3 – which is important on time and voltage scales relevant to integrated circuits.

Dielectric relaxation can have an important effect on the performance of circuits. Capacitors in sample-and-holds, switched capacitor filters, charge redistribution A/D converters and many other kinds of analog circuits are used to sample the value of a changing voltage waveform at periodic intervals. Charge which becomes trapped in the dielectric during a sampling period cannot readjust instantaneously when a new

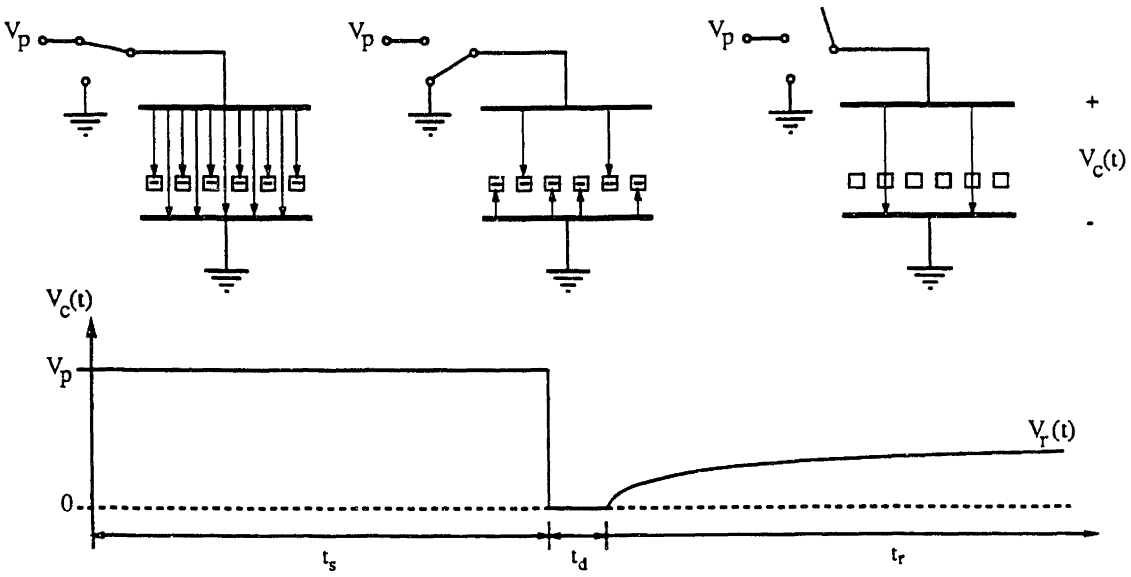


Figure 1.3: Dielectric relaxation in a capacitor due to a distribution of traps close to the bottom plate (a) Charged to voltage V_p for time t_s, during which electrons are trapped from the bottom plate (b) Discharged to ground for time t_d, insufficient for all traps to discharge (c) Plates are open-circuited and a recovery voltage $V_r(t)$ develops as trapped charge is emitted to the bottom plate.

voltage is sampled, causing errors to build up and propogate from one clock cycle to the next [17]. Since these errors depend on the past history of the input, they are very difficult to eliminate by circuit techniques.

Although dielectric relaxation has always been recognized as an important source of error in discrete capacitor structures, experimental data on integrated circuit capacitors is scarce and the physical mechanism is still poorly understood. In this thesis a charge trapping mechanism is proposed to explain both threshold voltage relaxation in MOSFETs and dielectric relaxation in integrated circuit capacitors. A complete characterization of these phenomena and an understanding of their physical origins are prerequisites for their successful elimination or reduction by circuit, device or processing solutions.

1.2 Organization of the Thesis

This thesis is divided into two main parts; the first, consisting of Chapters 2–5, is devoted to theory while the second, comprising Chapters 6 and 7 discusses measurements and experimental results.

Chapter 2 begins with a review of various polarization mechanisms and instabili-

ties which occur in the Si-SiO_2 system. In addition to charge trapping, there are many other physical mechanisms such as mobile ions, dipolar and interfacial polarizations and thermal effects, which under certain conditions, can manifest themselves in similar ways. It is shown that these contributions can be neglected at the frequencies of interest in circuit applications and it is therefore concluded that the relaxation effects are due to the tunneling of charge to and from oxide traps.

Threshold voltage relaxation in MOSFETs and dielectric relaxation in capacitors are just two particular manifestations of the tunneling of charge between the interface and oxide traps under large-signal stress conditions. Oxide charge trapping can also reveal itself in other ways, depending on experimental conditions and the particular variables which are chosen as the outputs in a measurement. For example, the emission of charge from oxide traps can cause a slow decay of the current in charge pumping measurements [18] or a long capacitance transient in DLTS measurements [19]. Section 2.2 looks at some of these previous examples in which oxide traps have been observed to affect MOS device characteristics and measurements.

Without loss of generality, the remainder of this thesis will focus only on threshold voltage relaxation in MOSFETs and dielectric relaxation in capacitors. It is sufficient to consider only these two effects for several reasons: (1) they are of fundamental importance in circuit applications, (2) there is a scarcity of experimental data and a lack of adequate theoretical models explaining these effects and (3) any theory which describes the physics underlying these two effects is also applicable to any other manifestation of charge trapping by slow states.

This thesis addresses a range of levels of modeling, beginning with a detailed analysis based on the fundamental physics, moving through a series of approximate solutions and concluding with some very simple circuit models. Chapter 3 begins with a discussion of the theory of tunneling to and from oxide traps. Various channels through which tunneling may occur are presented in Sec. 3.2 and it is shown that, by careful design of the measurement conditions, all but the two most probable of these can be neglected. These two channels are elastic tunneling transitions between oxide traps and either interface states or band states in the semiconductor. Time constants and rate equations are derived for these two processes in Secs. 3.3–3.5.

Chapter 4 extends the theory to include the effect of band bending in the oxide, by which it becomes possible for an oxide trap to charge and discharge by different mechanisms. A new formalism is introduced which takes into account all possible combinations of initial and final states in elastic tunneling. Approximate expressions for the voltage and time dependence are derived in Sec. 4.4. The remainder of Chapter 4 considers the theoretical temperature and geometry dependence of the threshold voltage shifts and discusses possible hysteresis effects. Chapter 5 then applies the theory developed in Chapters 3 and 4 to dielectric relaxation in capacitors and derives exact and approximate relations for the time and voltage dependence of the recovery voltage.

The experimental portion of the thesis begins in Chapter 6 with a discussion of the issues and problems involved in the measurement of the small relaxation effects. A simple closed-loop measurement circuit is presented in Sec. 6.2.2 which enables the isolation of the trapping-induced threshold voltage shift from other possible contributions due to thermal gradients, hot carrier effects and other component nonidealities. Experimental results obtained by these measurement techniques are presented in Chapter 7 for conventional and nitrided SiO_2 gate dielectrics, together with simulation results based on the theory of Chapters 4 and 5. The relation of threshold voltage relaxation to $1/f$ noise is discussed and a simple circuit model is presented with which the effects of relaxation on circuit performance can be simulated using SPICE. Section 7.3 presents measurements and simulations of dielectric relaxation in capacitors as a function of bias conditions, voltage and temperature.

Finally, Chapter 8 concludes with a summary of the major results and some suggested directions for future research in this area.

Chapter 2

Background

2.1 Polarization Mechanisms

An ideal linear, time invariant capacitor has a unique charge for each value of the applied voltage, independent of past history. Any real dielectric used in MIS devices, however, does not respond instantaneously to a change in voltage. Dielectric relaxation refers to the time delay associated with the various polarization processes in the insulator in response to a changing electric field.

When a DC voltage is suddenly applied across an insulating material, the movement of charge can be classified into four contributions:

- DC Conductivity
  An insulator can pass a steady flow of current which is controlled by the potential barriers at the contacts as well as by any space charge contained within the dielectric. If carriers have sufficient energy to surmount the potential barrier or a sufficient voltage is applied for tunneling throught the barrier to occur, then carriers can be injected into the conduction band of the insulator. In addition to this electronic component, conduction may occur by the migration of ions in materials such as glass, or due to mobile ionic contaminants. The injection and trapping of electrons in the insulator or the buildup of ions at the electrodes can lead to a buildup of uncompensated space-charge in the dielectric which modifies the potential. Under these conditions, if the barriers are not limiting, the current can become space-charge limited [20], leading to relaxation effects as discussed in Sec. 5.3.

- Electronic Polarization
  The electronic component arises from the displacement of the electron cloud relative to the nucleus. These displacements are spatially limited and have time constants in the optical frequency range.

- Ionic Polarization
  The ionic contribution is caused by displacements of charged ions towards the electrodes which causes a change in interionic spacing. This polarization forms more slowly than the electronic contribution, but is still essentially instantaneous at electrical frequencies. The electronic and ionic polarization therefore contribute a nearly constant amount to the dielectric constant at the frequencies of interest.

- Anomalous (Absorptive) Polarization
  The remaining polarization mechanisms are of the slowly forming or "absorptive" type, with time constants in the electrical frequency range. These mechanisms result in a slowly relaxing current component when when a DC voltage is applied or removed from an insulator and under AC excitations, contribute to dielectric losses and dispersion in the dielectric constant over a limited range of frequencies.

It is the slowly forming polarizations of the latter group which are of interest in this thesis. Since we are interested in effects relating to circuit performance, only a small subset of all possible relaxation mechanisms are of importance. In particular, we only need to consider those mechanisms which can respond to a stimulus on time scales comparable to normal circuit operation times, but which relax over a much longer period of time when the stimulus is removed. The stimulus is either directly, or indirectly, due to the application of a voltage across the dielectric of a MOSFET or a capacitor, and the relevant time scales depend on the particular circuit configuration and the required accuracy. For example, in a MOSFET input stage of an amplifier or comparator, the only relevant events are those which can be initiated in the time interval during which the input may be subjected to large-signal voltage excursions, but which leave persisting effects over time scales comparable to the settling time of the amplifier. For a capacitor in a data converter or a switched-capacitor filter, the mechanisms of interest are those which respond during a clock period, e.g. 1-10 μs, but which require a greater amount of time to relax, leaving effects which persist through several clock cycles. Because the long relaxation times of these mechanisms leave residual effects which persist for periods longer than the time it took to initiate them, they often manifest themselves as hysteretic effects in circuits. Some of the mechanisms which are candidates for explaining the relaxation effects in MOSFETs and capacitors are discussed below.

2.1.1 Mobile Ions

It is well known that the drift of ionic contaminants such as sodium can shift the flatband voltage in CV measurements of MOS capacitors [21, 22]. The time constant for drift of positively charge ions across 1000$\AA$ of SiO_2 are on the order of 100-1000

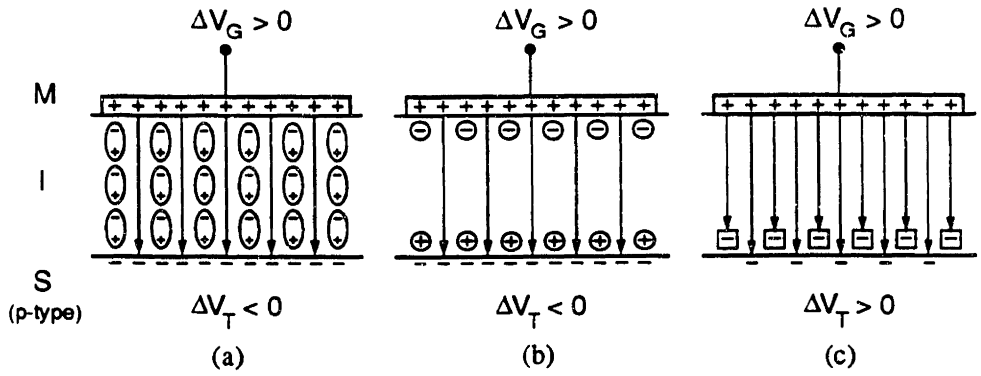


Figure 2.1: (a) Dipolar polarization (b) mobile ions (c) trapping by interface states or near-interface oxide traps

seconds or higher at room temperature [22]. Thus, the time required to drift across a typical tunneling distance of 10$\AA$ [Sec. 2.1.4] would be on the order of 1-10 seconds, many orders of magnitude greater than the tunneling time constant (Section 2.1.4), and time scales of interest in typical circuit applications. In response to a positive gate voltage stress, ionic drift causes the buildup of positive charge at the insulator-semiconductor interface as shown in Fig. 2.1(b). Thus, $\Delta V_T/\Delta V_G < 0$, which is the opposite polarity of the observed threshold voltage shift.

2.1.2 Dipolar/Interfacial Polarizations

The dipolar polarizability arises from molecules with a permanent dipole moment. The theory as presented by Debye [23] postulates a viscous damping effect on polar molecules due to molecular interactions, which affects their response time in a changing electric field. Interfacial polarizations refer to the accumulation of charge at structural interfaces in heterogeneous materials. Polarization in a two layer dielectric having different conductances and dielectric constants was first proposed by Maxwell [24]. Wagner [25] later analyzed the case of conductive spheres embedded in a dielectric medium. Other examples of interfacial polarizations include the buildup of charge at grain boundaries in polycrystalline materials, mobile ions and perhaps, the exchange of charge by tunneling across an interface. These slow forming, or "absorptive" polarizations are capable of storing energy in the dielectric which is not immediately available when the drive is removed.

While data on SiO_2 is scarce, Snow [26] has analyzed a polarization effect in phosphosilicate glass films which is believed to be an interfacial polarization due to small amounts of a conductive phase, perhaps a crystalline precipitate, embedded uniformly in the glass. The approach to saturation shows a broad distribution of time constants, but the effective time constant, defined as the time required for ΔV to come within $1/e$ of its final value, is on the order of 100 minutes at room temperature, far

too slow to cause errors at typical analog circuit frequencies. Rossel et al [27] also observed threshold voltage shifts in MOSFETs due to dipolar polarizations which were shown to be localized in a passivating layer of phosphorus glass rather than in the SiO_2.

In the dielectrics studied in this thesis, phosphorus glasses are not utilized at any point in the processing of the SiO_2 so that dipoles should not contribute to the measured hysteresis. More importantly, as shown in Fig. 2.1(a), for dipolar polarizations, $\Delta V_T/\Delta V_G < 0$, the wrong polarity to explain the observed shifts. Like mobile ions, these instabilities have largely been eliminated by clean processing conditions in modern technologies.

2.1.3 Fast Surface States

Fast surface states, or interface states, reside close enough to the Si-SiO_2 interface that the wavefunction of a trapped electron overlaps the wavefunctions of channel charge carriers and communication with the silicon bands can occur by a thermal (phonon-assisted) mechanism rather than by tunneling. For a positive gate voltage, charge trapping in fast surface states results in layer of negative charge along the Si-SiO_2 interface, as shown in Fig. 2.1(c). Thus, $\Delta V_T/\Delta V_G > 0$, the correct polarity to explain the observed shifts. As first shown by Shockley, Read and Hall [28] and reviewed in Sec. 3.5.2, the filling and emptying of fast surface states by channel charge carriers obeys a first-order differential equation governed by the time constant

$$\tau_{it}(E) = \frac{1}{\sigma_{no}\overline{v}(n_s + n_1) + \sigma_{po}\overline{v}(p_s + p_1)} \tag{2.1}$$

where,

$$\begin{aligned} n_1(E) &= n_s \exp[(E - E_F)/kT] \\ &\approx N_c \exp[-(E_c - E)/kT] \qquad (E_C - E_F > 3kT) \\ p_1(E) &= p_s \exp[(E_F - E)/kT] \\ &\approx N_v \exp[-(E - E_v)/kT] \qquad (E_C - E_F > 3kT) \end{aligned}$$

where N_c and N_v are the effective densities of states in the conduction and valence bands, σ_{no} and σ_{po} are capture cross sections for electrons and holes and $\overline{v}$ is the thermal velocity. Consider three special cases, assuming an NMOS transistor and some typical values for σ_{no}, σ_{po} and n_s:

1. Strong Inversion, $(n_s \gg n_1)$

$$\begin{aligned} \tau_{it}(E) &\approx \frac{1}{\sigma_{no}\overline{v}n_s} \\ &\approx \frac{1}{(10^{-15}\mathrm{cm}^2)(10^7\mathrm{cm/s})(10^{18}\mathrm{cm}^{-3})} \approx 10^{-10}\mathrm{s}. \end{aligned}$$

2. Accumulation, ($p_s \gg p_1$)

$$\tau_{it}(E) \approx \frac{1}{\sigma_{po}\bar{v}p_s} \approx \frac{1}{(10^{-15}\text{cm}^2)(10^7\text{cm/s})(10^{18}\text{cm}^{-3})} \approx 10^{-10}\text{s}.$$

3. Instrinsic $n_s = p_s = n_i$
(also assume $E = E_F$, $\sigma_{po} = \sigma_{no} = \sigma$)

$$\tau_{it}(E) = \frac{1}{4\sigma\bar{v}n_i} \approx \frac{1}{4(10^{-15}\text{cm}^2)(10^7\text{cm/s})(10^{10}\text{cm}^{-3})} \approx 2.5\text{ms}.$$

Since MOSFETs in analog circuits are normally biased in strong inversion, the response of the interface states is typically quite fast, on the order of 100ps. Thus, under normal bias conditions the fast surface states respond too quickly to explain the long relaxation times (> 100ms) associated with the threshold voltage hysteresis. However, when the surface carrier concentration is low, as in depletion or weak inversion, the time constant for interface states can become long and strongly temperature dependent.

2.1.4 Oxide Traps

The occupancy of oxide traps also satisfies a first-order differential equation with a time constant that increases exponentially with depth from the interface, as discussed in Sec. 3.3. Assuming a rectangular tunneling barrier of height ϕ_c=3eV between the Si and SiO_2 conduction bands, the time constant is,

$$\tau_{ot}(x) = \tau_o \exp(x/x_o) \quad (2.2)$$

$$x_o = \hbar/2\sqrt{2m\phi_c} \approx 1\mathring{A} \quad (2.3)$$

$$\Longrightarrow \tau_{ot} \approx 160\mu s. \qquad \text{for } x{=}10\ \mathring{A}$$

where $\tau_o \approx 10^{-10}s$ [29] is the time constant at the interface. The exponential dependence of the time constant provides a time constant dispersion spanning many orders of magnitude which can easily explain circuit errors on the time scales of interest. For example, Eq. (2.2) yields a time constant of about 2μs for a trap located $10\mathring{A}$ and 50ms for a trap located $20\mathring{A}$ from the interface. As shown in Fig. 2.1(c), oxide traps produce threshold shifts with the correct polarity ($\Delta V_T/\Delta V_G > 0$) to explain the experimental observations.

2.1.5 Defect Relaxation

We have assumed that the energy of a trap remains constant after accepting or emitting an electron. However, it is possible for the energy of a trap to change depending on its charge state. An electron may tunnel into an unrelaxed (empty) defect configuration at an energy E_t which then decreases by an amount ΔE. The electron then tunnels out of the relaxed (filled) state at an energy $E_t - \Delta E$. Zvanut et al [30] measured such a relaxable defect with a relaxation energy $\Delta E \approx 1eV$.

Breed [31] has proposed two types of traps in SiO_2, both of which have ground states below the top of the Si valence band. The excited (empty) state of the type 1 trap, with $\Delta E_1 \approx 0.6eV$ falls within the Si bandgap, while the excited state for the type 2 trap with $\Delta E_2 \approx 1.3eV$ is above the Si conduction band.

Defect relaxation causes hysteresis effects in MOS CV measurements and cannot be ruled out as a source of threshold voltage hysteresis in MOSFETs. However, these effects are typically observed under much higher voltages than those used in typical analog circuits.

In addition to these relaxation mechanisms, which are properties of the dielectric and the silicon-insulator interface, the following effects may also contribute to measurements of threshold voltage hysteresis:

2.1.6 Hot Carrier Effects

It is well known that the hot carrier effect can cause an increase in the threshold voltage of an NMOS transitor, due to the creation of interface states and the trapping of charge in the oxide [32, 33, 34, 35, 36, 37]. While the origins of hot carrier degradation are still incompletely understood, three distinct types of damage have been identified: interface state generation, oxide trapped charge and "relaxable" traps. Oxide trapped charge is primarily responsible for the degradation in NMOS devices at high gate bias, $V_{GS} \approx V_{DS}$, where hot electron injection into the oxide is at a maximum [38]. The creation of interface states is enhanced by the injection of both electrons and holes into the oxide [39, 40], a condition which peaks at $V_{GS} \approx V_{DS}/2$, which is where the maximum threshold voltage shifts occur. These types of damage are of a permanent nature.

The third type of damage, the so-called "relaxable" states, are thought to be created during hot hole injection, which is maximized around $V_{GS} \approx V_{DS}/4$. These states appear to be neutral when created but can be filled and emptied by the injection of electrons or holes into the oxide by channel hot-carrier stressing or by the application of a large bias to the gate ($\pm$4MV/cm) [41, 35]. Following stress at $V_{GS} = V_{DS}/2$, Doyle [42] has reported that some of the created damage dissappears with a time constant $\tau_r \approx 600 - 900$s when the device is grounded at low V_{DS}. This relaxation was originally thought to be due to an annealing of the damage [42], but is now believed

to be due to the tunneling emission of carriers trapped in defects created during stress [41].

The creation of interface states and relaxable traps may be enhanced under dynamic stress conditions, where the drain voltage is held constant at a high value and a square wave is applied to the gate [40, 43], since both electrons and holes will then be injected into the oxide.

These effects are of great interest since they may contribute to the observed threshold voltage shifts under certain bias conditions, especially in saturation, which is the normal bias region for MOSFETs in analog circuits. However, it is important to first understand the threshold voltage relaxation due to the exchange of charge with existing traps, in the absence of any hot carrier effects. The creation of defects, especially of the relaxable variety, must be eliminated from the measurement, as they will interfere with the effect we are trying to isolate. For this reason, the threshold voltage relaxation measurements must be carried out in the linear region with 50mV $< V_{DS} <$ 100mV, where impact ionization and the associated hot carrier damage is negligible [44]. The difference between dynamic hot carrier stressing (high V_{DS}) and our pulsed measurement conditions (low V_{DS}) should be emphasized. In the former, threshold voltage relaxation effects are caused by the capture of electrons by oxide traps created during stress, whereas in the latter, the shifts are due to the charging of pre-existing traps.

2.1.7 Thermal Hysteresis

The circuit described in Section 6.2.2 actually measures the change in drain current due to an applied gate voltage pulse. The change in drain current depends on both mobility μ and threshold voltage V_T. In saturation, the dominant term in the temperature dependence of the drain current is μ, which follows a $T^{-3/2}$ power law, causing the drain current to decrease with temperature. In the subthreshold region the V_T term dominates and the current increases with temperature. Cuevas [45] has demonstrated that temperature changes due to changes in device power dissipation may be indistinguishable from trapping relaxation effects. It is therefore of great importance to ensure that the power dissipation through the device under test be maintained at a constant level so that this source of hysteresis may be ruled out.

Assuming that hot carrier and thermal effects can be eliminated by proper design of the measurement apparatus, we can conclude that, at the time scales of interest, threshold voltage relaxation is probably due primarily to tunneling of charge between the interface and pre-existing oxide traps, with the possible contribution, under certain temperature and bias conditions, of the fast surface states. While defect relaxation may potentially contribute to the hysteresis at high stress voltages, we neglect this effect for now.

2.2 Previous Studies

Surface states at the Si-SiO_2 interface introduce allowed energy levels within the silicon bandgap which can exchange charge with the substrate, causing degradation in the performance of MOS devices. These trapping centers are traditionally divided into two classes: the fast states which are located at the interface plane and slow states located within the oxide. The interface (or fast) states are in intimate electrical contact with the substrate and communicate readily with the silicon conduction and valence bands by a thermal process, as modeled by Schockley, Read and Hall [28]. These traps are characterized by a single time constant at any given energy. Charge transfer between the oxide traps (slow states) and the substrate occurs either by overcoming the high potential barrier presented by the oxide, or by tunneling through the oxide layer, leading to a very wide range of time constants. In this section we survey previous studies of the effects of oxide traps on MOS device characteristics and measurements.

2.2.1 Small-Signal Measurements

1/f Noise

The effects of oxide traps on the small-signal performance of MOSFETs have been studied extensively. McWhorter showed that the tunneling mechanism provided the time constant dispersion necessary to explain the $1/f$ spectrum of flicker noise in germanium filaments [2]. Later, this theory was verified experimentally and extended to explain the $1/f$ noise in MOSFETs [46, 29, 47, 48, 49, 50]. The actual noise spectrum is not strictly $1/f$, but goes as $1/f^{\gamma}$, where $0.7 < \gamma < 1.2$. The exponent γ has been shown to depend on gate voltage through the shift in oxide trap energies due to the electric field in the oxide, an effect which also influences the threshold voltage hysteresis, as shown in Section 4.3.1. Measurements of $1/f$ noise have also been investigated as a technique to measure oxide trap densities [51, 52, 53, 54].

Conductance and Capacitance-Voltage (C-V) Measurements

The standard techniques for studying surface states at the semiconductor-insulator interface utilize capacitance-voltage and conductance measurements. The accepted model for interpreting the results of these experiments treats all traps as though they are located directly at the Si-SiO_2 interface where they communicate with the semiconductor bands by a thermal mechanism. However, the parallel conductance of a MOS structure often shows a dispersion in time constants greater than that predicted by the standard Shockley-Read-Hall analysis of trapping by interface states. Some authors have explained this effect by the tunneling of charge to and from oxide traps [55, 56], while others have shown that an accurate model requires the superposition

of contributions due to both the oxide traps and the interface states [57].

Heiman and Warfield [58] studied the effects of oxide traps on MOS capacitance measurements and noted that, under certain bias conditions there may be an asymmetry in the depth to which traps fill and empty, leading to hysteresis in CV curves. The model which they developed forms the basis for most analyses of the effects of oxide traps on MOS device characteristics and will be discussed in more detail in Section 2.3.

2.2.2 Large-Signal Measurements

The focus of this thesis is on the effects of oxide traps on the large-signal transient response of MOS circuits. In contrast to small-signal measurements, where fluctuations in trap occupancy in a narrow energy window a few kT wide around the Fermi level are of interest, in large-signal measurements the Fermi level may sweep across a wide range of trap energies at the interface and band bending in the oxide may be substantial. While studies of the effects of oxide traps under large-signal conditions have been limited, several areas in which they impose important limitations on the performance and measurement of MIS devices are reviewed below.

Threshold Voltage Relaxation in SiO_2

Charge trapping hysteresis in MOSFETs was studied by Lundstrom et al [59, 60] who measured drain current drifts in MOSFETs corresponding to threshold voltage shifts of 100mV-200mV. They observed that the shift was a logarithmic function of time, a result which they modeled by an analysis paralleling that of Heiman and Warfield, assuming a continuous distribution of oxide traps in energy. There are difficulties with this analysis, as will be discussed in Sec. 2.3. With stress times on the order of minutes to hours and voltages ranging from 7-25 V, their measurement conditions were far outside the normal operating range of modern analog integrated circuits.

Rossel et al [61] observed slow drifts in drain current and threshold voltage following a step in gate voltage of sufficient amplitude to invert the surface. Following a decrease in drain current during a long stress period persisting for many hours, it was found that the current could be restored to its original value by the application of a short accumulation pulse, on the order of seconds. This behavior was explained by the slow emission of majority carriers from a discrete trap level, located close to the energy level of the substrate dopant impurities, during stress, and the faster capture of majority carriers during cutoff. The origin of the traps was believed to be due to the spreading of dopant impurities into the oxide during thermal growth. Under normal circuit bias conditions, however, such a majority carrier trap always lies well above or below the Fermi level and its occupancy will remain unchanged unless the device is driven close to cutoff. Therefore, this model cannot explain hysteresis effects in the bias regions of interest in most circuits and particularly, in the bias regions

studied in this thesis (Chapter 6). The smallest stress times in Rossel's study were on the order of tens of seconds, which are too long to be of interest in most analog circuit applications.

In previous work on threshold voltage hysteresis, NMOS transistors subjected to positive polarity gate voltage pulses of 1-5V on time scales ranging from 1μs to 10ms exhibited threshold voltage shifts from 100μV to 1mV [62, 10]. PMOS transistors showed substantially less hysteresis under the same stress conditions. Negative stress voltages produced no measurable hysteresis in either channel type. A strong correlation was found between threshold voltage hysteresis and 1/f noise, in support of the tunneling model.

Charge Pumping

Charge pumping is a technique for measuring surface state densities on MOS transistors or capacitors [63, 64, 65, 66, 67, 68, 69, 70, 71]. A voltage pulse is applied to the gate, alternately filling surface states with holes and electrons and causing a recombination current to flow through the substrate connection. A saturation level is expected for the charge pumping current when the gate pulse is sufficient to drive the Fermi level all the way to one the band edges, at which point all of the traps will be filled with either electrons or holes. However, the charge pumping current is observed to increase slowly with voltage above the saturation value. One component of this excess current depends logarithmically on the duration of the gate voltage pulse and it has been suggested that this is the result of the tunneling of charge into oxide traps [18].

Deep-level Transient Spectroscopy

In deep-level transient spectroscopy (DLTS) the interface states are filled during a pulse and the emission rates are analyzed during a temperature scan. Again, the DLTS spectrum should saturate for sufficiently large pulse amplitudes because all the interface states are filled. Instead, the amplitude of the DLTS spectrum is observed to increase when the filling pulse amplitude or the pulse width increases. This effect has also been explained by the tunneling of charge carriers between oxide traps and the semiconductor surface[72, 73].

Threshold Voltage Instabilities in III-V and II-VI MISFETs

Threshold voltage instabilities in many III-V and II-VI compound semiconductors pose a serious problem in the attempt to realize electronic devices. The drain current of a MISFET is observed to decrease as a function of time after the application of a positive gate voltage step which induces an accumulation of electrons in the channel. This drain current drift may persist for minutes to hours. Several mechanisms have

been proposed to explain this drain current drift phenomenon, including nearest-neighbor hopping of anion vacancies [74, 75] and the capture and emission of free carriers by defects located in the near interface region [76]. In the disorder-induced gap state (DIGS) model [77, 78] these defects are distributed energetically in the bandgap of both the insulator and semiconductor, and spatially in the insulator layer near the interface. Interface state densities in these compounds are much higher than in silicon and in GaAs may be so high as to prevent inversion by pinning of the Fermi Level near midgap.

Koelmans and de Graaf [79] analyzed this drift in thin film CdSe FETs and found the decay to be approximately logarithmic with time. Many researchers have reported that the drain current drift is strongly temperature dependent, disappearing at low temperatures. Based on Heiman and Warfield's model, they explained the effect due to the tunneling of carriers into two discrete trap levels in the insulator, one falling energetically within the bandgap of the semiconductor and one above the bottom of the conduction band. Electrons are captured by thermally assisted tunneling into the higher energy trap and by temperature independent tunneling into lower one.

Drain current drifts in InP MISFETs have also been explained by the tunneling of charge into discrete trap levels in the insulator [80, 81, 82]. It has also been suggested by Goodnick [83] that the tunneling occurs, not to discrete trap levels, but to a conducting bulk-like layer of In_2O_3 within the native oxide at the interface. More recently, Van Staa and Rombach [84] have conducted constant capacitance deep level transient spectroscopy (CC-DLTS) studies on InP-SiO_2 MIS capacitors and were able to explain the capture and emission characteristics using Heiman and Warfield's model with a continuous spatial and energetical distribution of traps within the bandgap. While most studies confirm a logarithmic time dependence, Prasad [85, 86] showed that the time dependence in InP MISFETs is closer to $\sqrt{\Delta V} \propto \log(t)$, a result that was explained for a monoenergetic trap level by including oxide band bending in the Heiman and Warfield analysis. Some authors report a temperature dependence of the drift similar to that observed by Koelmans and de Graaf in CdSe [80] while others [87] contend that the temperature dependence is merely a measurement artifact. Drain current drifts have also been observed in InSb transistors and have been explained by the same basic mechanisms [88].

Clearly, this problem is still only partially understood, and an investigation of threshold voltage hysteresis in the Si-SiO_2 system may contribute to an understanding of the drain current drift in compound semiconductors.

Dual-Dielectric Memory Devices

A hysteresis effect also occurs in the flat-band voltage of MNOS structures [89, 90], an effect which has been exploited in the fabrication of memory devices. In this case, the traps of interest are usually located in a second dielectric, separated from the Si substrate by a layer of SiO_2. Under sufficiently high voltages, traps in the

second layer can be lowered to the Si valence band energy where they are filled by tunneling electrons. When the stress is removed the charge remains trapped, until the application of the opposite polarity voltage exceeds a threshold voltage where the trapped carriers line up in energy with the Si conduction band and can tunnel back out to empty states. The SiO_2 layer must be sufficiently thick to prevent leakage to fast surface states at the interface. Defect relaxation [30], as discussed in the next section, may also contribute to hysteresis effects in dual-dielectric memory devices. We shall see later that the conduction band-to-valence band (cb/vb) and valence band-to-conduction band (vb/cb) tunneling processes do not normally occur in the voltage ranges of interest in analog circuits.

2.3 The Model of Heiman and Warfield

While Heiman and Warfield [58] were primarily concerned with deriving the contribution of oxide traps to the small signal capacitance of the MOS structure, they noted the possibility of hysteresis effects in CV curves and their method of analysis has formed the basis for most analyses of the effect of oxide traps on MOS device characteristics. The model has a simple intuitive appeal but several difficulties which are worth mentioning briefly.

The basic approach consists of first deriving the capture cross section of an oxide trap as seen by an electron at the interface [58]:

$$\sigma_n(E,x) = \sigma_{no}(E)\exp(-x/x_o) \quad (2.4)$$

$$\sigma_p(E,x) = \sigma_{po}(E)\exp(-x/x_o) \quad (2.5)$$

where x_o, as defined in Eq. 2.3, is a characteristic length on the order of $1\mathring{A}$ and $\sigma_{no}(E)$ and $\sigma_{po}(E)$ are the capture cross sections for electrons and holes, respectively, by an interface state at energy E. This result follows in an almost obvious way from the exponential decay of the electron wavefunction with depth from the interface.

The change in trapped carrier concentration at energy E and depth x is then calculated by a Shockley-Read-Hall (SRH) analysis [28], with a time constant obtained by substituting the effective capture cross section, Eq. (2.4) into Eq. (2.1):

$$\begin{aligned}\tau_{ot}(E,x) &= \frac{1}{\sigma_{no}\overline{v}(n_s+n_1)+\sigma_{po}\overline{v}(p_s+p_1)}\exp(x/x_o) \quad (2.6)\\ &= \tau_o\exp(x/x_o)\end{aligned}$$

where τ_o is the time constant at the interface ($x=0$).

There are questions as to the validity of a direct application of the SRH analysis to tunneling transitions between electrons at the interface and traps in the oxide. There are several reasons for this; (1) The SRH analysis was derived for thermal transitions between a band edge electron or hole and a trap at the same general location in space.

The SRH kinetic picture, in which an electron moving at the thermal velocity will be captured if it comes within the capture radius of a trap, does not even make sense in the case of tunneling, where there is no physical current flowing in the conduction band of the insulator. (2) If tunneling is elastic, then the capture rate is not proportional to the total electron concentration as in SRH theory, but depends only on those electrons at the same energy as the trap level. This is especially important for the accurate modeling of the temperature dependence of trapping, as discussed in Sec. 4.5. (3) For a one-dimensional barrier potential, the transmission probability for an electron impinging on the interface depends on the component of its wavevector parallel to the interface. This makes sense intuitively, since an electron with all of its momentum directed parallel to the interface will certainly have a smaller probability of tunneling into an oxide trap than an electron with all of its momentum directed perpendicular to the interface. As shown in Sec. B.1, the tunneling probability decreases exponentially as the component of the wavevector parallel to the interface, $\vec{k}_{||} = (0, k_y, k_z)$, increases. Thus, tunneling is dominated by the density of a one-dimensional subset of states with $\vec{k}_{||} \approx 0$.

This method treats the oxide traps formally as if they were interface states communicating thermally with the conduction band, with the tunneling process taken into account implicitly by an effective capture cross section which increases exponentially with depth into the oxide. Although the derivation is based on the tunneling of a conduction band electron into an oxide trap at constant energy, the result has been applied by Heiman and Warfield and others [58, 59, 79, 84, 80] to the tunneling of carriers into oxide traps with energies falling within the silicon bandgap. There is no explicit reference in Heiman and Warfield's original paper as to the physical mechanism by which traps in the band gap might be filled, however the SRH analysis implicitly assumes a thermal mechanism.

While Heiman and Warfield's approach has been successful in explaining many of the qualitative features of data, it does not rest on a sound theoretical foundation and fails to describe the more detailed voltage, time and temperature dependence of the data, as presented in Chap. 7. The derivation also assumes Boltzmann statistics, which is not valid under high gate biases when the Fermi level is near or in the conduction band. Unfortunately, these are the conditions of interest in our pulsed gate voltage experiments and in many other studies to which this type of analysis has been applied.

In Sec. 3.4.3 we shall present a specific mechanism which does lead to a time constant of the same general form as Eq. 2.6. The new derivation yields increased insight into the actual process taking place and provides an interpretation and method for calculating σ_{no} and σ_{po} in Eq. 2.4 quantum mechanically. We shall also investigate other mechanisms for dissipating energy during the transition, which lead to different forms of the tunneling time constant.

Chapter 3

Theory of Tunneling to and from Oxide Traps

3.1 Introduction

In modeling semiconductor device behavior for the purpose of understanding or predicting its effect on circuit performance, two general attitudes can be taken. The first, or empirical, approach seeks a mathematical model or analog which simulates chosen aspects of the observed behavior, without demanding a physical understanding of the underlying mechanisms. Dow's model for dielectric relaxation in a capacitor, discussed in Chapter 5, is an example of this kind of model. Success is measured by some criterion of fit of the model to the data, without requiring any connection between the underlying physics and the parameters needed to achieve such fit. This approach may be demanded by lack of more detailed understanding or may be motivated, even when an exact description is possible, by the requirements of computational speed and efficiency in a circuit simulation environment.

The second approach seeks, above all, an understanding of the underlying physics. This physically-based approach starts from fundamental mechanisms, deriving the system behavior and comparing with experiment. Success is based on the ability of the model to predict or explain the results of future experiments.

The first approach seeks to describe, the second to explain, the phenomenon of interest and both approaches have their place in the overall modeling effort. The orientation of this thesis is primarily on the second approach. This is based on the belief that, while the prediction and improvement of the performance of devices in circuits is our ultimate objective, an understanding of the mechanisms responsible for threshold voltage relaxation is a prerequisite for the development of effective processing, device and circuit level solutions to the problem. Armed with a qualitative and quantitative physical understanding, the development of fast and efficient models for circuit-level simulators then becomes a relatively straightforward task.

The first step in attempting to formulate the problem is to identify and isolate the dominant mechanisms responsible for the observed behavior. This chapter is devoted to a theoretical description of the fundamental mechanisms which cause the threshold voltage relaxation effects. This theory is then applied in Chapters 4 and 5 to predict the response of MOSFETs and capacitors under pulsed gate voltage stress. These results are verified against experimental data in Chapter 7. In Sec. 7.2.6 we will use what we have learned to construct a simplified model for threshold voltage relaxation which exemplifies the first approach to modeling; capturing the essential features of the observed behavior while bypassing the detailed physics.

3.2 Fundamental Mechanisms

In Chapter 2 it was shown that the polarity and long relaxation times of the observed threshold voltage shifts could only be explained by a mechanism involving the tunneling of charge to and from oxide traps. The mechanisms for charging and discharging of oxide traps in an MIS structure can be classified into two groups, (1) the capture (emission), by traps in the insulator, of electrons or holes from (to) the insulator conduction and valence bands [Fig. 3.1(a)], and (2) the direct tunneling of electrons and holes between the metal or semiconductor and traps in the insulator [Fig. 3.1(b)].

3.2.1 Insulator Band-to-Trap Transitions

Processes in the first group have been studied extensively, both for their importance as long-term degradation mechanisms in MOS devices [91] and as a tool for studying the properties of insulating barriers [92, 93]. The capture process requires two steps; (1) the injection of electrons or holes into the insulator from the metal or semiconductor, or their generation within the insulator itself and (2) the capture of these injected carriers by a trap in the insulator.

Once an electron or hole is present in the oxide conduction or valence band, the capture step may occur through various energy-loss mechanisms, including thermal (phonon-assisted), radiative (photon-assisted) or Auger-impact (particle, i.e. electron- or hole-assisted) processes. Since the latter two are comparatively weak [94, 92, 95], the capture process is generally assumed to be thermal and is described by SRH kinetics [28]. Thus, the rate at which traps are filled is given by [92, 96]

$$\frac{dn_{ot}}{dt} = \sigma_n f(D_{ot} - n_{ot}) - \nu n_{ot}$$

where $f = n\bar{v}$ is the flux and $\nu = N_c\bar{v}\exp[-(E_c - E_t)/kT]$ is the emission rate. It should be emphasized that, despite its widespread and general use, this expression is strictly valid only for the thermal capture of electrons propagating through the conduction or valence bands of the insulator.

The first, or injection step, may be accomplished by:

1. Generation of carriers within the oxide, e.g. by internal photoemission (IPE), kilo-volt electron beam generation (keV-EBG), X-ray or vacuum ultra-violet light electron injection (VUVEI).

2. Fowler-Nordheim tunneling electron injection (FN-TEI).

3. By imparting to the carriers in the semiconductor sufficient energy to surmount the semiconductor-insulator potential barrier, e.g. by avalanche electron injection (AEI).

For MOS devices in circuits under normal operating conditions, only the second two mechanisms, shown schematically in Fig. 3.1(a), need be considered. Fowler-Nordheim tunneling occurs when the electric field is high enough to lower the oxide conduction band sufficiently close to the semiconductor conduction band for electrons to tunnel through the triangular barrier. In this regime, the tunneling current is a strong function of the electric field $\mathcal{E}$ across the oxide [97],

$$J = CF^2 e^{-\beta/F} \tag{3.1}$$

where C and β are functions of the barrier height and the effective masses [98, 99]. For the oxide thicknesses and fields used in MOSFETs in analog IC processes, Fowler-Nordheim tunneling is usually negligible under normal circuit operating conditions. For example, with an oxide thickness of $500\mathring{A}$, 15V across the oxide would decrease the barrier thickness to $100\mathring{A}$ at the silicon conduction band edge, which is still too thick for substantial tunneling currents. Fowler-Nordheim tunneling will generally be insignificant for fields less than about 7 MV/cm [92], a condition which is satisfied in all of our measurements and in virtually all analog circuits.

Since the oxide barrier is too high for significant thermal injection, the only way carriers can gain sufficient energy to surmount the barrier is through photons (internal photoemission) or by acquiring energy through impact ionization. In circuits, only the latter mechanism may contribute. Substrate hot carrier injection occurs when the silicon is driven into deep depletion by the application of a high gate voltage. Electrons and holes generated by avalanche multiplication in the semiconductor space charge layer are accelerated toward the interface by the high electric field with sufficient energy to surmount the barrier. In MOSFETs, this is generally not of any importance, since gate voltages are low and the presence of the source and drain regions prevents the substrate from deep-depleting.

In analog circuits however, MOSFETs are are usually biased in saturation and the generation of hot carriers by impact ionization in the high lateral fields at the drain end of the channel is a problem which cannot be completely avoided. While the contributions of channel hot carriers cannot be overlooked, it will be shown in Chapter

7 that threshold voltage relaxation effects occur even in the linear region, at voltages too low for the generation of hot carriers. Furthermore, it is found that the threshold voltage shifts are a very weak function of drain voltage, so that even at moderate levels of saturation, hot carriers contribute negligibly. The technique used to measure the threshold voltage relaxation, as described in Chapter 6, is designed to exclude all mechanisms from the first group. It can therefore be concluded that only the second category, consisting of direct tunneling transitions between the semiconductor surface and oxide traps, can be responsible for the observed threshold voltage shifts.

3.2.2 Tunneling to and from Oxide Traps

While processes from the first category have been researched extensively, particularly by the digital community, because of their long term effect on device reliability, the effect of the direct tunneling mechanisms on device performance has received very limited attention [1, 91]. It will be shown in this thesis that, in addition to the long term stability issues, the direct tunneling of charge carriers to and from oxide traps affects the immediate, short-term operation of MOS devices and consequently, poses a potentially even more important limiting factor on the performance of precision and high-speed analog circuits. This group of mechanisms, involving tunneling transitions to and from traps, can be classified according to the nature of (1) transitions at the interface, (2) transitions in the barrier and (3) transitions within the oxide trap:

1. Transitions at the Interface

At the Si-SiO_2 interface, three types of initial states are available; conduction band states, valence band states and interface (fast surface) states. Since interface states communicate thermally with the conduction and valence bands, any oxide trap charging/discharging process which includes these states is inherently a multi-step process. The first step in this process is the thermal capture of a carrier from the silicon conduction or valence band into an interface state, which may occur by either a cascade process or lattice relaxation multiphonon emission [100], as discussed in (3) below. The details of the capture process are not important, since it is well known that the interface states can be modeled within the SRH theory. The second step is the tunneling of the electron from an interface state into an oxide trap. The process may also proceed in the reverse direction.

A two-step process of this type, involving tunneling through a real intermediate state (trap-assisted tunneling), was first proposed by Sah [101, 102] to explain the excess currents in Esaki diodes. In this case, tunneling between the conduction and valence bands occurred through a deep bulk level. Fu and Sah [46] later proposed two step-tunneling through interface states to explain 1/f noise in MOSFETs.

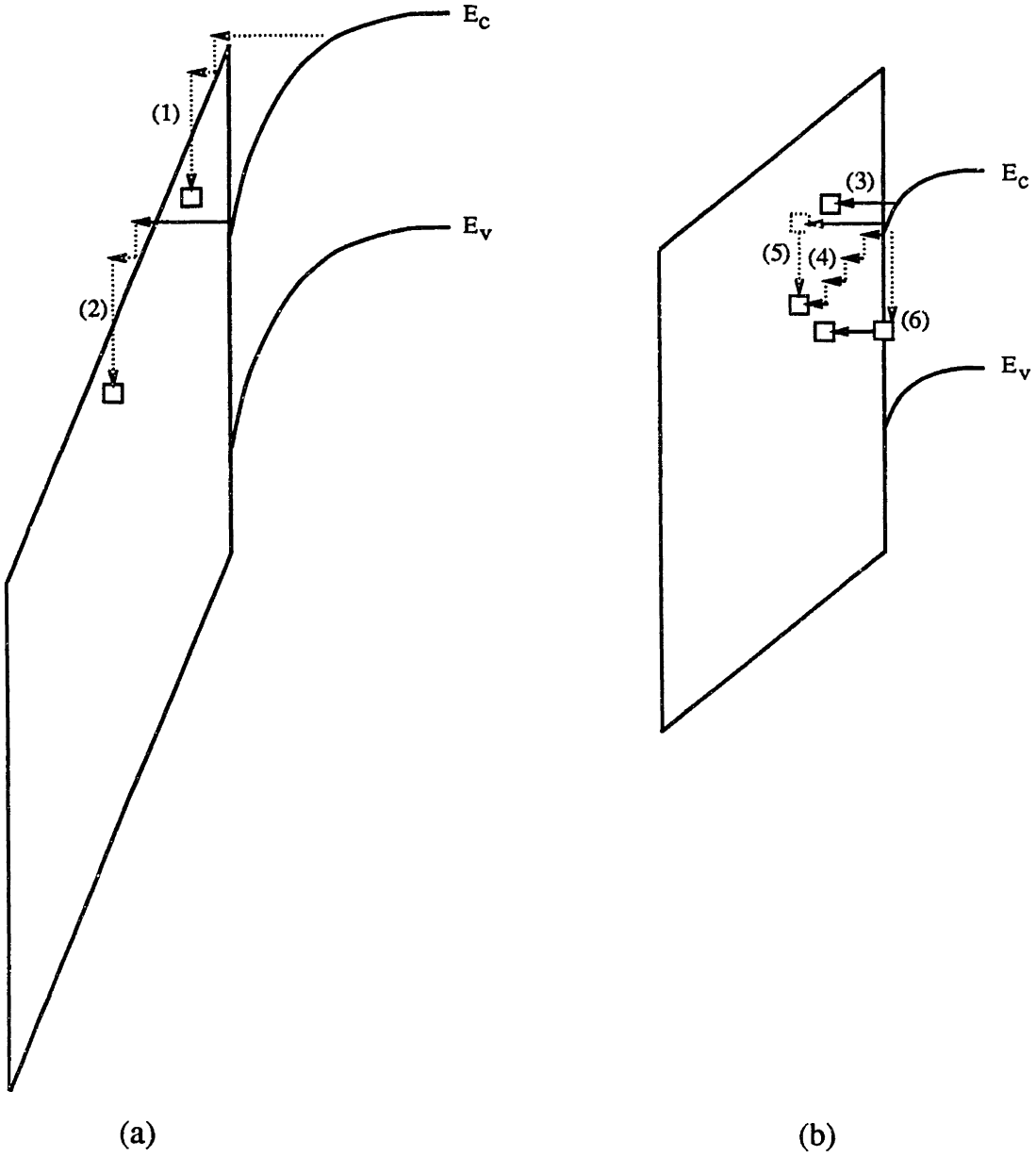


Figure 3.1: Mechanisms for introducing charge carriers into SiO_2. (a) Thermal capture of electrons injected into the insulator conduction band; (1) avalanche electron injection (AEI), (2) Fowler-Nordheim tunneling injection (FNTEI), (b) Direct tunneling to and from oxide traps; (3) elastic tunneling from the conduction band, (4) inelastic (phonon-assisted) tunneling, (5) elastic tunneling followed by cascade phonon emission or lattice relaxation multiphonon emission, (6) two-step, interface state-assisted tunneling.

2. Transitions in the Barrier

The transition through the barrier region occurs by tunneling, since the measurement, by design, excludes the possibility of injection of carriers into the oxide bands. Tunneling can occur from an occupied initial state to an unoccupied final state through either elastic or inelastic channels:

A. Elastic Tunneling Channel

For a time-independent barrier Hamiltonian, tunneling is inherently an elastic process (Eq. A.12), requiring the presence of occupied states on one side of the barrier and unoccupied states at the same energy on the other side, a fact that has been overlooked in much of the literature [58]. Tunneling can occur directly from the silicon conduction or valence bands into or out of oxide traps in a narrow energy window around the band edges where free carriers are available, or by the two-step (trap-assisted) process described above. Since the interface states are continuously distributed in energy throughout the silicon bandgap [103, 104], the elastic tunneling channel can contribute to the charging and discharging of oxide traps over a wide range of energies throughout the silicon bandgap.

B. Inelastic tunneling channel

If the barrier Hamiltonian includes terms that depend explicitly on time, it is possible for the electron or hole to gain or lose energy in the process of tunneling. The time dependent barrier fluctuations which stimulate the inelastic tunneling and provide the required change in energy may take the form of various elementary excitations in the barrier. These may be quasi-particle excitations, such as electrons, holes, excitons and photons, or collective excitations, such as phonons, plasmons, and magnons.

Phonon-assisted inelastic tunneling has been widely studied and is known to cause discontinuities in the slope of silicon and germanium tunnel diode current-voltage characteristics [105] at small forward and reverse voltages corresponding to the emission of phonons required to conserve momentum in tunneling across the indirect bandgap.

3. Transitions at the Trap Site

The final state is, by assumption, a bound energy level of an oxide trap potential well. However, the final state into which the carrier tunnels need not necessarily be the equilibrium ground state of the defect. We discuss below two possible mechanisms which are essentially the reverse of the trap-assisted process discussed above, in which capture occurred by a thermal transition to an interface state, followed by

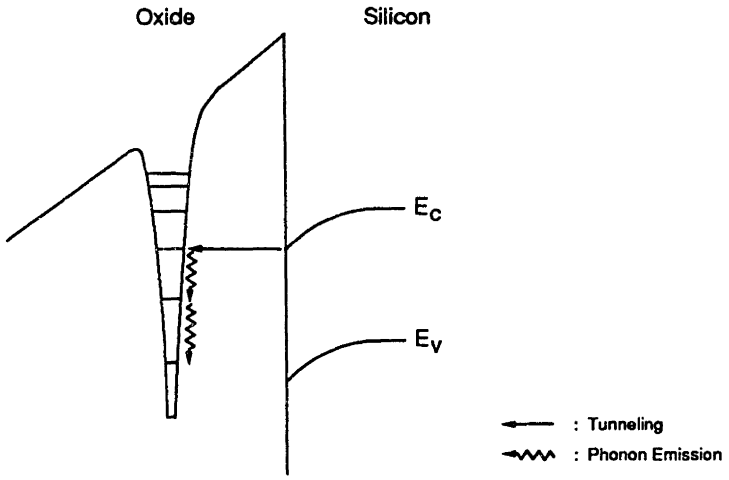


Figure 3.2: Elastic tunneling followed by cascade phonon emission.

a tunneling transition. We now consider the case where tunneling occurs into an excited state of a trap, followed by a thermal transition into the ground state. There are two mechanisms by which this thermal step might occur. These mechanisms were originally proposed to explain the microscopic mechanisms responsible for SRH thermal recombination events in bulk semiconductors, but the principles apply equally well to transitions at the interface or within the oxide, the only difference being that, in the latter, the first step is a tunneling transition.

A. Cascade Phonon Emission

This process is thought to be the dominant capture mechanism in shallow defects with coulombic attractive potentials, which have a number of bound state energy levels analogous to the hydrogen atom. An example of such a level in SiO_2 is the 2.4 eV trap thought to be associated with immobile sodium [92]. A conduction band electron that comes within the large capture radius of an excited state, or an electron which tunnels directly into an excited state as shown in Fig. 3.2, can lose energy in steps by cascading down the ladder of excited levels to the ground state [106, 95, 107, 108, 109].

The cascade process was originally proposed by Lax [106, 95] to explain the very large capture cross sections of a wide variety of traps in Si and Ge. Since many of these traps have binding energies of many times the Debye energy, the question arose of how the large energy change is dissipated during a capture event. As we have seen, a single transition to the ground state by multiphonon emission is an improbable process, and could not explain the large cross sections measured. Provided that the separation between excited states is less than the Debye energy kT_D, each of the

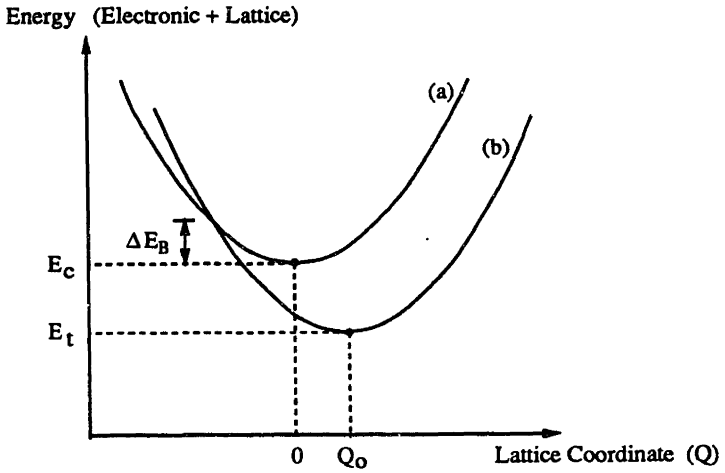


Figure 3.3: Configuration coordinate diagram showing (a) system consisting of an electron in the conduction band and an empty oxide trap (b) system consisting of an electron in a bound state of an oxide trap.

transitions in the cascade process requires the emission of only one phonon and can occur readily. However, there is also a finite probability that the electron will be thermally reemitted back to the conduction band or a higher excited state during any of these intermediate stages, leading to a capture cross section which increases with decreasing temperature. The rate limiting step is the last transition, from the first excited state to the ground state, which may require a multiphonon transition. However, the carrier is effectively trapped, having been removed from the conduction band and undergone a decrease in energy in the process, regardless of whether it actually enters the ground state or not.

B. Lattice Relaxation Multiphonon Emission

The majority of traps in SiO_2 are neutral and are believed to capture electrons and holes by lattice relaxation multiphonon emission (MPE) [92, 110, 111, 112, 113]. Fig. 3.3 shows a configuration-coordinate diagram of the changes in the total energy of the system which occur during the capture process. Initially, the system consists of an empty defect in the oxide and an electron at the interface, which we assume is in the conduction band, but may also be in an interface state. The total energy of the system consisting of electron + lattice is just

$$E = E_c + AQ^2$$

where E_c is the conduction band energy and the elastic energy of the lattice is a quadratic function of the configuration coordinate. In this case there is minimal coupling between the delocalized electron and the defect and the energy is minimized

when the lattice is in its undistorted equilibrium configuration with $Q = 0$, as shown in Fig. 3.3(a).

After capture, the presence of the electron at the defect site changes the bonding of the lattice ions with their nearest neighbors. The lattice becomes polarized in the vicinity of the electron and the equilibrium positions of the ions change. To first order, the lattice deformation linearly modulates the electron energy,

$$\delta E_e = -BQ$$

where $B = -\partial E_e/\partial Q$ is the deformation potential [114]. The total energy of the system consisting of the trapped electron and the surrounding lattice is then

$$\begin{aligned} E &= AQ^2 - BQ \\ &= -AQ_o^2 + A(Q - Q_o)^2 \end{aligned}$$

where the new energy minimum is at $Q_o = B/2A$. Thus, the presence of the electron on the defect site lowers the equilibrium energy and shifts the coordinate as shown in Fig. 3.3(b). This is analogous to the formation of a small polaron [115].

Capture occurs when, for sufficiently large lattice vibrations, which increase with temperature, the two curves cross. At this point the electronic energies differ but the total energies are equal and there is strong mixing between the initial and final state wavefunctions. A tunneling transition can then occur from the interface into the trap. Since the lattice cannot readjust as quickly as the electronic transition occurs (Franck-Condon principle), immediately after capture the lattice will be in a highly excited, nonequilibrium state due to the distortion created by the electron. Relaxation to the final equilibrium state, $Q = Q_o$, occurs by dissipating the excess energy through multiple phonon emission. Since there is an energy barrier ΔE_B for capture, the process is thermally activated and the cross section increases with temperature.

The lattice relaxation mechanism has received strong support from measurements of low-frequency $1/f$ noise due to single defect states in MOSFETs [116, 117, 118] and by measurements of the shift between optical and thermal ionization energies in Si-SiO_2 interface states [119]. It has also been reported in measurements of the temperature dependence of the flatband voltage shift in avalanche injection experiments on SiO_2 [120] and in two-level current fluctuations in metal–silicon-dioxide–silicon tunnel diodes [121].

3.3 Tunneling Time Constant

In this section we derive a general differential equation describing tunneling transitions between an oxide trap at energy E_r on the right side of a potential barrier, and a set of states at energies E_l on the left side of the barrier, as shown in Fig. 3.4.

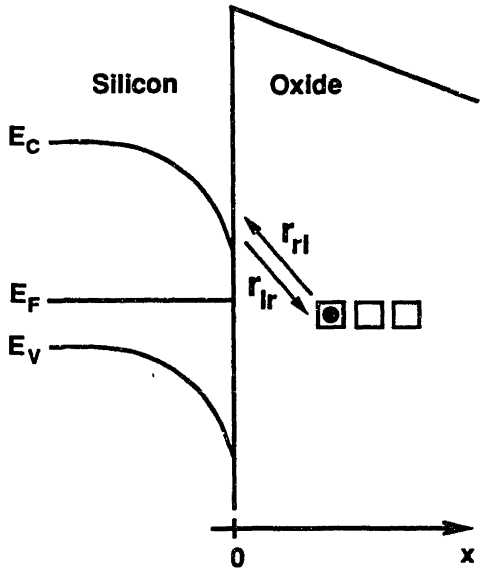


Figure 3.4: Band diagram showing transitions between states at energy E_l at the interface and states at energy E_r in the oxide

It is assumed that states on the left and on the right side of the barrier are in equilibrium among themselves so that they can be described by Fermi-Dirac statistics, with quasi-Fermi levels F_l and F_r respectively,

$$f_l(E_l) = [1 + \exp(E_l - F_l)/kT]^{-1} \tag{3.2}$$

$$f_r(E_r) = [1 + \exp(E_r - F_r)/kT]^{-1} \tag{3.3}$$

In general, the transition rate ($s^{-1}cm^{-3}\ eV^{-1}$) from states on the left with energies in the range E_l to $E_l + \Delta E_l$ to states on the right at energy E_r is given by

$$r_{lr}(E_l, E_r, \vec{k}_{||}) = W_{lr}(E_l, E_r, \vec{k}_{||}) f_l(E_l) \rho_l(E_l)[1 - f_r(E_r)]\rho_r(E_r)\Delta E_l \tag{3.4}$$

where ρ_l (eV^{-1}) and ρ_r ($cm^{-3}\ eV^{-1}$) are the densities of states on the left and right sides of the barrier, respectively. $W_{lr}(E_l, E_r, \vec{k}_{||})$ is the probability per unit time of an electron making a transition from a state at energy E_l to a state at energy E_r in the barrier. In general this probability is also a function of $\vec{k}_{||}$ because the WKB transmission probability for a one-dimensional barrier depends on the component of the electron wavevector k_x normal to the interface [Eqs. B.12 and B.13].

The rate of the reverse process is

$$r_{rl}(E_r, E_l, \vec{k}_{||}) = W_{rl}(E_r, E_l, \vec{k}_{||}) f_r(E_r) \rho_r(E_r)[1 - f_l(E_l)]\rho_l(E_l)\Delta E_l \tag{3.5}$$

The transition probability W_{rl} is related to W_{lr} by detailed balance. In equilibrium, $f_l = f_r = f$ and $F_l = F_r = E_F$ where f is the equilibrium Fermi function. The

meaning of equilibrium in the insulator will be addressed further in Sec. 3.5.3. The condition that the two rates self-balance in equilibrium yields

$$W_{rl} = W_{lr} \exp[(E_r - E_l)/kT] \tag{3.6}$$

The net rate $r(E_r)$ at which electrons flow into states at energy E_r is obtained by taking the difference between Eqs. 3.4 and 3.5, summing over all values of $\vec{k}_{||}$ and integrating over energy:

$$\begin{aligned} r(E_r) &= \sum_{\vec{k}_{||}} \int [r_{lr} - r_{rl}] dE_l \\ &= \sum_{\vec{k}_{||}} \int W_{lr}(E_l, E_r, \vec{k}_{||}) \rho_l(E_l) \rho_r(E_r) \\ &\qquad \times \left\{ f_l(E_l)[1 - f_r(E_r)] - e^{(E_r - E_l)/kT} f_r(E_r)[1 - f_l(E_l)] \right\} dE_l \\ &= \frac{\partial}{\partial t}[f_r(E_r)\rho_r(E_r)] \end{aligned} \tag{3.7}$$

Cancelling $\rho_r(E_r)$ from both sides, since it is independent of E_l, and noting that

$$1 - f_l(E_l) = \exp[(E_l - F_l)/kT] f_l(E_l)$$

Equation 3.7 can be rewritten in the form

$$\frac{\partial f_r}{\partial t} + f_r[1 + \exp[(E_r - F_l)/kT] \sum_{\vec{k}_{||}} \int W_{lr} \rho_l f_l dE_l = \sum_{\vec{k}_{||}} \int W_{lr} \rho_l f_l dE_l \tag{3.8}$$

The time constant can be identified as

$$\frac{1}{\tau(E_r, x)} = \{1 + \exp[(E_r - F_l)/kT]\} \sum_{\vec{k}_{||}} \int W_{lr}(E_l, E_r, \vec{k}_{||}) \rho_l(E_l) f_l(E_l) dE_l \tag{3.9}$$

The right side of Eq. 3.8 is just $\{1 + \exp[(E_r - F_l)/kT]\}^{-1} \tau^{-1} = f_l(E_r)/\tau$. Thus, the differential equation governing the occupancy of oxide traps is

$$\frac{\partial f_r(E_r)}{\partial t} + \frac{f_r(E_r)}{\tau} = \frac{f_l(E_r)}{\tau} \tag{3.10}$$

In the following two sections the time constant, Eq. 3.9, is evaluated for several special cases of interest. Two possible types of initial states must be considered; band states in the silicon and fast surface states at the interface.

3.4 Tunneling between Oxide Traps and Band States

In this section we consider the case where the initial state is a state in the conduction or valence band of the semiconductor. For simplicity we evaluate only the case of the conduction band, the valence band being analogous.

In this case, ρ_r is the density of oxide traps and ρ_l is the one-dimensional density of conduction band states at fixed $\vec{k}_{||}$, including spin,

$$\rho_l = \frac{m^* L_x}{\pi \hbar^2 k_x} \quad (\mathrm{eV}^{-1}) \tag{3.11}$$

$$\rho_r = D_{ot}(E,x) \quad (\mathrm{cm}^{-3}\ \mathrm{eV}^{-1}) \tag{3.12}$$

where m^* is the effective mass in the semiconductor, assumed isotropic, and L_x is the extent of the semiconductor perpendicular to the interface. Since $k_x^2 + \vec{k}_{||}^{\,2} = k^2 = 2m^*E/\hbar^2$ is fixed, ρ_l is a function of $\vec{k}_{||}$. We assume all energies to be measured upward from the silicon conduction band.

The density of states Eq. 3.11 assumes that electrons in the silicon can be treated as a free Fermi gas, as in the Sommerfeld model of a metal. In fact, the electrons in the inversion or accumulation layers form a two-dimensional electron gas, confined to a narrow potential well at the interface, with quantized energy levels [122, 123, 124]. At the relatively high temperatures and low fields of interest we shall neglect this quantization, consistent with the bulk of semiconductor physics on which MOS device equations are based [104].

The occupancy functions on the left and right sides of the barrier [Eqs. 3.2 and 3.3], in this case, are just the corresponding functions for electrons at the interface and traps in the oxide,

$$\begin{aligned} f_l &= f(E,t) \\ f_r &= f_{ot}(E,x,t) \end{aligned} \tag{3.13}$$

In general, both of these depend on time.

Equation 3.10 for the oxide trap non-steady-state occupancy function becomes

$$f_{ot}(E,x,t) = f(E) + [f_0(E) - f(E)]e^{-t/\tau_{ot/cb}}$$

where f_0 and f are the equilibrium Fermi functions at the interface before stress ($t = 0^-$) and a very long time after stress ($t \to \infty$).

Subsituting Eq. 3.11–3.13 into Eq. 3.9 and converting the sum over $\vec{k}_{||}$ to an integration using,

$$\sum_{\vec{k}_{||}} \to \frac{L_y L_z}{(2\pi)^2} \int 2\pi \vec{k}_{||} d\vec{k}_{||} \tag{3.14}$$

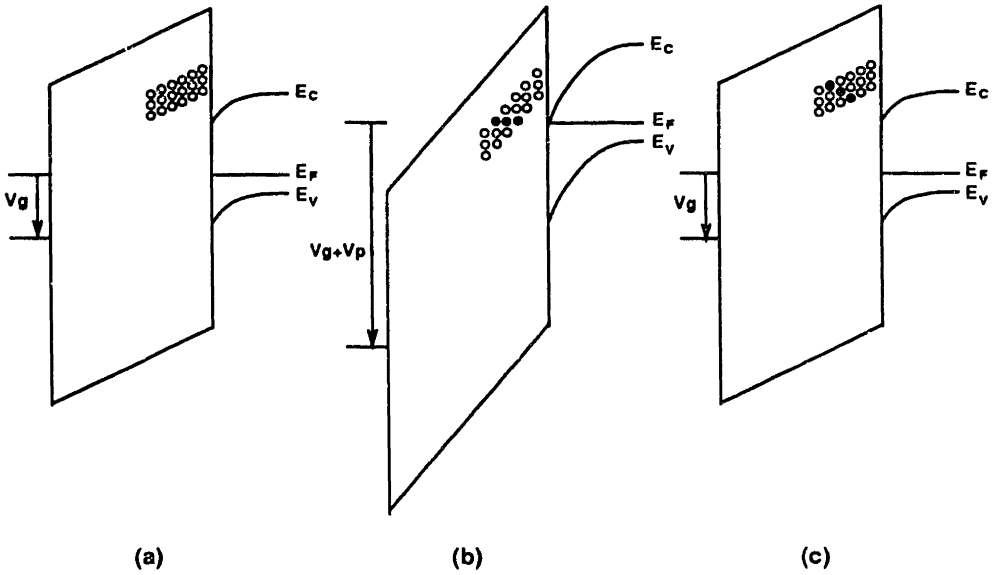


Figure 3.5: Elastic tunneling from the conduction band. (a) Before stress; traps empty, (b) During stress; traps fill by tunneling from filled conduction band states, (c) After stress; electrons tunnel out to empty conduction band states.

where $\Omega = L_x L_y L_z$ is the volume of the semiconductor, the time constant can be written,

$$\frac{1}{\tau_{ot/cb}(E_r, x)} = \frac{\Omega m^*}{2\pi^2\hbar^2}\left[1 + e^{(E_r - F_l)/kT}\right]\int_0^\infty dE_l \int_0^{\vec{k}_{||_{max}}} d\vec{k}_{||}\, \vec{k}_{||} f_l \frac{W_{lr}}{k_x} \tag{3.15}$$

where $\vec{k}_{||_{max}} = (2m^*E/\hbar^2)^{1/2}$ and the energy integration is over all conduction band states, with the zero reference for energy taken at the bottom of the conduction band. In a quantized description of the inversion layer, the continuous energy integration would be replaced by a summation over subbands.

The value of the tunneling rate W_{lr} and hence the time constant, depend on whether the final state is at the same energy as the intial state and, if not, on the mechanism by which the energy is dissipated. We explicitly evaluate Eq. 3.15 for several cases below.

3.4.1 Elastic Tunneling

The band diagram of an NMOS transistor before stress is shown in Fig. 3.5(a), with a distribution of traps in the oxide indicated by open circles. When a positive stress voltage is applied to the gate, the oxide bands bend as shown in Fig. 3.5(b), lowering some of the trap energies to within tunneling range of the conduction band. Traps

with energies in a window a few kT wide around the conduction band can then be filled by tunneling. Note that traps below the Fermi level require some mechanism for energy loss, as will be discussed in the following sections, but cannot be filled by elastic tunneling from the conduction band. When the stress is removed, these trap energies rise again, relative to the interface, and the trapped electrons can be reemitted to empty conduction band states as shown in Fig. 3.5(c). Of course, this mechanism applies equally well to the tunneling of electrons at energies below the valence band.

The transition probability is given by Fermi's Golden Rule [Eq. A.2],

$$W_{lr}(E_l, E_r, \vec{k}_{||}) = \frac{2\pi}{\hbar}|M_{lr}(E_l, \vec{k}_{||})|^2 \delta(E_l - E_r) \tag{3.16}$$

where M_{lr} is Bardeen's effective matrix element for tunneling, given by Eq. A.8. Substituting this expression into Eqs. 3.9 and 3.15, the time constant is

$$\begin{aligned} \frac{1}{\tau_{ot/cb}(E_r, x)} &= \frac{2\pi}{\hbar}\sum_{\vec{k}_{||}} |M_{lr}(E_r, \vec{k}_{||})|^2 \rho_l(E_r) \\ &= \frac{\Omega m^*}{\pi\hbar^3}\int_0^{\vec{k}_{||_{max}}} \frac{|M_{lr}|^2}{k_x}\vec{k}_{||}\, d\vec{k}_{||} \end{aligned} \tag{3.17}$$

The details of the evaluation of Eq. 3.17 are presented in Appendix B for two model trap potentials. Assuming a parabolic dispersion relation in the conduction band of the oxide, the result is

$$\begin{aligned} \tau_{ot/cb}(E, x) &= \tau_o(E, x)\exp\left[2\int_0^x K(x')dx'\right] \\ &= \tau_o(E, x)\exp\left[\frac{4}{3}\left(\frac{2M^*}{\hbar^2}\right)^{1/2}\frac{(\phi_c - E + q\mathcal{E}x)^{3/2} - (\phi_c - E)^{3/2}}{q\mathcal{E}}\right] \end{aligned} \tag{3.18}$$

where $\mathcal{E} = -V_{ox}/t_{ox}$ is the electric field across the insulator. A two-band model for K will be investigated in Sec. 3.6.

The value of the prefactor $\tau_o(E, x)$ depends on the model used for the trap potential. Assuming a one-dimensional delta function for the oxide trap, from Eq. B.32,

$$\begin{aligned} \frac{1}{\tau_o(E, x)} &= \frac{q\sigma_n\mathcal{E}}{4\pi\hbar^2}\left(\frac{2m^{*2}}{M^*}\right)^{1/2}\left[\frac{\phi_c - E + q\mathcal{E}x}{(\phi_c - E + q\mathcal{E}x)^{1/2} - (\phi_c - E)^{1/2}}\right] \\ &\quad \times\left\{1 - \exp\left[-\frac{2E}{\hbar}\left(\frac{2m^{*2}}{M^*}\right)^{1/2}\frac{(\phi_c - E + q\mathcal{E}x)^{1/2} - (\phi_c - E)^{1/2}}{q\mathcal{E}}\right]\right\} \end{aligned} \tag{3.19}$$

where σ_n is the capture cross section of an oxide trap for an electron and m^* and M^* are the effective masses in the semiconductor and in the oxide respectively. If the

trap is modeled as a three dimensional spherical delta function, then, from Eq. B.72,

$$\frac{1}{\tau_o} \approx \frac{2\hbar}{m^* x} \frac{kK^2}{K^2 + k^2} \left[1 - \exp\left(\frac{-k^2 x}{K}\right)\right] \tag{3.20}$$

where k and K are defined in Appendix B. The one-dimensional and three-dimensional results are equal, in the case of a rectangular barrier, if we identify the "capture cross section" in Eq. 3.19 as

$$\sigma_n = 8\pi \left(\frac{M^*}{m^*}\right)^2 \frac{k}{K(k^2 + K^2)}$$

3.4.2 Inelastic Tunneling

In the case of inelastic tunneling, the transition rate is given, from Fermi's Golden Rule (Eq. A.2), by

$$W_{lr} = \frac{2\pi}{\hbar} |\langle \psi_r | H'_{el-ph} | \psi_l \rangle|^2$$

where H'_{el-ph} is the electron-phonon interaction term in the Hamiltonian.

In normal-mode coordinates Q_i, the interaction potential can be expanded in a Taylor series [125],

$$V = U_c + \sum_i V_i Q_i + \cdots \sum_{ijk\cdots n} V_{ijk\cdots n} Q_i Q_j Q_k \cdots Q_n + \cdots$$

where U_c is the static periodic potential, $V_i = \partial U_c / \partial Q_i$, $V_{ij} = \partial^2 U_c / \partial Q_i \partial Q_j$, etc. When written in second quantized form [115, 126, 127], each of the Q_i's either creates or annihilates one phonon. Thus, there are several ways in which two states ψ_a and ψ_b, differing in phonon occupation number can be coupled by this Hamiltonian. In first order perturbation theory, the nth order term, $V_{ijk\cdots n}$ in the expansion will couple states differing by n phonons, with the emission or absorption of n phonons. It is also possible for ψ_a and ψ_b to couple through the first order term $\sum_i V_i Q_i$ via virtual intermediate states in higher orders of perturbation theory. While the actual calculation of phonon-assisted tunneling probabilities is quite involved, it might be expected that, since these processes involve higher orders of perturbation theory or higher order anharmonic terms in the expansion of the electron-phonon interaction potential, the transition probability would decrease rapidly with increasing n for non-polar materials where the electron-phonon interaction is relatively weak. Indeed, it is well known that multiple phonon emission, even when the carrier and defect are located at the same point in space, is a highly improbable process [106, 95, 107], and one might expect it to be even more unlikely when the initial and final state wavefunctions are widely separated in space, as in phonon-assisted tunneling.

Another process for inelastic tunneling involves impurity molecules in the oxide. When the energy difference between the tunneling electrons and the oxide traps corresponds to the vibrational frequency of molecules in the oxide, $\Delta E = h\nu$, excess energy

can be dissipated through the excitation of vibrational modes of the molecules. These may arise from OH groups, present either chemically within the oxide, or as absorbed residual water vapor [128, 129]. Exposure to room air or other organic contaminants can introduce additional vibrational modes ranging in energy from 0.05 to 0.5eV.

In SiO_2 the highest optical phonon energy is 0.153 eV [130] so, if tunneling occured only from the conduction band, trap energies more than this amount below the band edge would require a multiphonon emission. A related problem where inelastic tunneling competes with a two-step process arises in conjuction with excess currents in MIM tunnel diodes [101, 102]. Kane [131] has calculated the excess currents without gap states, with energy and momentum conservation provided by photons, phonons, multiphonons, electron-electron interactions and Auger processes. All of these yield results too small to explain the experimentally observed currents. Thus, the presence of a real intermediate gap state was found to be essential [132]. The situation at the MOS interface is similar and it is therefore reasonable to assume that, for traps below the conduction band, if an elastic tunneling channel through interface states and a phonon-assisted tunneling channel exist in parallel, then the latter process can be neglected.

In the following two sections we describe two such processes whereby an elastic tunneling step is accompanied by an energy loss through the occurence of a thermal transition, either at the defect site or at the interface.

3.4.3 Lattice Relaxation Multiphonon Emission (LRME)

We now consider the possibility that the energy of the defect site may relax upon capturing an electron, as described in Sec. 3.2.2. Consider the system consisting of trap + electron in the ground state, at some energy $E_r < E_l$. In order for the electron to tunnel back out to an empty conduction band state, the system must acquire an amount of energy $\hbar\omega = E_l - E_r + \Delta E_B$ from thermal vibrations of the lattice, where $E_l \approx E_c$ and ΔE_B is the barrier height in Fig. 3.3. The probability of the defect gaining this amount of thermal energy is proportional to the number of phonons $N(\omega)$ of energy $\hbar\omega$, which is given by the Bose-Einstein distribution

$$N(\omega) = \frac{1}{e^{\hbar\omega/kT} - 1} \tag{3.21}$$

For $\hbar\omega \gg kT$, we can approximate Eq. 3.21 by a Boltzmann factor, $\exp[-(E_l - E_r + \Delta E_B)/kT]$. However, the energy $\hbar\omega$ that can be supplied by a single phonon cannot be greater than the maximum energy in the phonon spectrum, approximately 0.153 meV in SiO_2 [130]. For energies greater than this, a multiple phonon process is required.

The multi-phonon problem can be approached from a statistical thermodynamic point of view. The system consisting of electron + defect is in equilibrium with

a reservoir (the lattice) at constant temperature and can therefore be treated in the canonical ensemble. The probability of the system being in a state of energy $E_l - E_r + \Delta E_B$ above the ground state is given by the Boltzmann distribution,

$$\frac{1}{kT}\exp[-(E_l - E_r + \Delta E_B)/kT]$$

where $1/kT$ (eV^{-1}) is a normalization factor. The transition rate from the defect to the conduction band is therefore given by the product of two probabilities,

$$W_{rl}(E_r, E_l, \vec{k}_{||}) = \left\{\frac{1}{kT}e^{-(E_l - E_r + \Delta E_B)/kT}\right\}\left\{\frac{2\pi}{\hbar}|M_{lr}(E_l, \vec{k}_{||})|^2\right\}$$

The first factor is the probability of the defect + electron system gaining sufficient energy to coincide with an allowed conduction band state in the silicon [Fig. 3.3]. The second factor is the probability of tunneling (elastically) at this energy. Once the electron has tunneled from the trap into the silicon, it may find itself in an excited state which relaxes back to the equilibrium state at energy E_l.

Using the detailed balance condition [Eq. 3.6], the rate of the reverse process is

$$W_{lr}(E_l, E_r, \vec{k}_{||}) = \frac{2\pi}{\hbar kT}|M_{lr}(E_l, \vec{k}_{||})|^2 e^{-\Delta E_B/kT} \tag{3.22}$$

Note that the tunneling matrix elements in both direction are the same because of the time reversal invariance of the Hamiltonian. This result means that, provided the electron has sufficient thermal energy to surmount the barrier in Fig. 3.3, it can tunnel elastically into an allowed vibrational state of the defect + electron system. Since the lattice cannot respond as fast as the electronic transition (Franck-Condon Principle), the defect immediately after the tunneling event is in a distorted nonequilibrium configuration. This excited state relaxes back to the ground state by multiphonon emission.

Substituting Eq. 3.22 into Eq. 3.9, the time constant for an oxide trap with ground state energy E_r and activation energy ΔE_B is

$$\frac{1}{\tau_{ot/lr}(E_r, x)} = \frac{2\pi}{\hbar kT}\left[1 + e^{(E_r - F_l)/kT}\right]e^{-\Delta E_B/kT}\sum_{\vec{k}_{||}}\int_{E_c}^{\infty} dE_l |M_{lr}(E_l, \vec{k}_{||})|^2 \rho_l(E_l) f_l(E_l) \tag{3.23}$$

Before attempting to evaluate this rather formidable integral, we make a simplification in order to gain some insight. Assume that M_{lr} varies slowly with E_l and is independent of $\vec{k}_{||}$, i.e., regarding all inversion layer electrons as though they can tunnel with equal probability, independent of their direction of travel. Then the one-dimensional density of states ρ_l can be replaced by the 3-d density $g_c(E_l)$ of conduction band states

$$g_c(E) = \frac{m_c^*\sqrt{2m_c^*(E - E_c)}}{\pi^2\hbar^3} \tag{3.24}$$

Eq. 3.23 can then be rewritten as

$$\frac{1}{\tau_{ot/lr}(E,x)} = \frac{2\pi}{\hbar kT} e^{-\Delta E_B/kT}\left[1 + e^{(E_r - F_l)/kT}\right]|M_{lr}|^2 \int_{E_c}^{\infty} g_c(E_l) f_l(E_l) dE_l$$

The integral is just the total electron concentration n_s. Thus the time constant can be written

$$\tau(E_r,x) = \left\{\frac{2\pi}{\hbar kT} e^{-\Delta E_B/kT}(n_s + n_1)|M_{lr}|^2\right\}^{-1} \tag{3.25}$$

where

$$n_1 = n_s e^{(E_r - F_l)/kT}$$

There are two important features of this time constant: (1) It is thermally activated and (2) it depends on inversely on n_s. The dependence on the total inversion charge concentration makes intuitive sense if a given trap is accessible to all conduction band electons. The time constant has the same form as that presented, by alternative derivations, in previous literature on lattice relaxation [117, 116],

$$\tau = \frac{e^{\Delta E_B/kT}}{\sigma_n \overline{v} n_s}$$

The only difference is that Eq. 3.25 also includes a tunneling step, as embodied in Bardeen's matrix element M_{lr}. It also contains the factor n_1 because we did not assume that the trap was initially empty, and included emission as well as capture in the formulation.

The probability of an electron tunneling actually depends on $\vec{k}_{||}$. Changing the summation over $\vec{k}_{||}$ to an integration using Eq. 3.14 and substituting for ρ_l from Eq. 3.11, Eq. 3.23 can be cast in the form

$$\frac{1}{\tau_{ot/lr}(E_r,x)} = \frac{1}{kT}\left\{1 + e^{(E_r - F_l)/kT}\right\} e^{-\Delta E_B/kT} \int_{E_c}^{\infty} f_l(E_l)\omega(E_l,x) dE_l$$

where $\omega(E_l,x)$ respresents the integral over $\vec{k}_{||}$ which was already carried out in Eq. B.1. The result, for a spherical delta function potential well is, from Eq. B.73,

$$\begin{aligned}\omega(E_l,x) &= \frac{\Omega m_l^*}{\pi\hbar^3}\int_0^{\vec{k}_{||_{max}}} \frac{|M_{lr}|^2}{k_x}\vec{k}_{||} d\vec{k}_{||} \\ &\approx \frac{2\hbar}{m_l x}\frac{kK^2}{K^2 + k^2} e^{-2Kx}\end{aligned} \tag{3.26}$$

Recognizing that $K^2 \gg k^2$, and making a Boltzmann approximation for f_l in order to carry out the energy integration,

$$\tau_{ot/lr}(E_r,x) = \tau_o(E_r) e^{\lambda_c(x)} \tag{3.27}$$

where the WKB integral $\lambda_c(x)$ is carried out at the conduction band energy, $E = 0$, where the tunneling step occurs,

$$\lambda_c(x) = \exp\left[\frac{4}{3}\left(\frac{2M^*}{\hbar^2}\right)^{1/2}\frac{(\phi_c + q\mathcal{E}x)^{3/2} - \phi_c^{3/2}}{q\mathcal{E}}\right] \tag{3.28}$$

and the prefactor is given by

$$\tau_o(E_r) = \frac{e^{\Delta E_B/kT}}{\sigma_n \overline{v}[n_s + n_1(E_r)]} \tag{3.29}$$

where $\overline{v} = 3kT/m^*$ (cm/s) and $\sigma_n = \hbar^3/24m^*\pi^2 x(kT)^2$ (cm^2). The time constant is of the same form as expected from Eq. 3.25.

The time constant [Eq. 3.29] can be cast into the Shockley-Read-Hall form [Eq. 2.1], with the oxide trap capture cross section identified as

$$\sigma_n(x,T) = \sigma_n e^{-\Delta E_{Bc}/kT}\, e^{-\lambda_c(x)} \tag{3.30}$$

where $\Delta E_{Bc} = \Delta E_B$ is the barrier for electron capture from the conduction band. In general, it is also possible for a hole to tunnel elastically from a valence band state into an oxide trap, followed by a lattice relaxation. By analogy with Eq. 3.30, the capture cross section for a hole is given by,

$$\sigma_p(x,T) = \sigma_p e^{-\Delta E_{Bv}/kT}\, e^{-\lambda_v(x)} \tag{3.31}$$

where E_{Bv} is the barrier for hole capture from the valence band and the WKB integral for holes is evaluated at the top of the silicon valence band

$$\lambda_v(x) = \frac{4}{3q\mathcal{E}}\left(\frac{2M_v^*}{\hbar^2}\right)^{1/2}\left[\phi_v^{3/2} - (\phi_v - q\mathcal{E}x)^{3/2}\right] \tag{3.32}$$

Using Eqs. 3.30 and 3.31 in the SRH time constant, Eq. 2.1, the time constant for the lattice relaxation process, including tunneling transistions from both bands can be written

$$\tau_{ot/lr} = \tau_o e^{\lambda_c(x)} \tag{3.33}$$

where

$$\tau_o = \left[\sigma_n(T)\overline{v}(n_s + n_1) + \sigma_p(T)\overline{v}(p_s + p_1)e^{\lambda_v(x)/\lambda_c(x)}\right]^{-1} \tag{3.34}$$

where the temperature dependence is now contained in the capture cross sections,

$$\sigma_n(T) = \sigma_n e^{-\Delta E_{Bc}/kT} \tag{3.35}$$

$$\sigma_p(T) = \sigma_p e^{-\Delta E_{Bv}/kT} \tag{3.36}$$

The precise values of σ_n and σ_p are not of interest and depend on the form of the potential assumed for the oxide trap. A more rigorous derivation of Eqs. 3.33 and 3.34 is presented in Appendix C.

In the case when $\Delta E_{Bc} = \Delta E_{Bv} = 0$, the time constant given by Eqs. 3.33 and 3.34 resembles the Heiman and Warfield result [Eq. 2.6], where the capture cross sections decrease exponentially with depth into the oxide. This provides a justification for treating traps in the oxide using a Shockley-Read-Hall analysis in the case of lattice relaxation multiphonon emission. The more rigorous derivation given here provides an explicit formula for calculating the values of the capture cross sections σ_n and σ_p from the tunneling matrix elements.

3.5 Tunneling between Oxide Traps and Interface States

The second mechanism for the charging of oxide traps is a two-step process in which interface traps serve as intermediate states. Electrons tunnel between oxide traps and interface states, which in turn, exchange charge with the silicon conduction and valence bands by a thermal process. This process provides another mechanism by which oxide traps with energies within the silicon bandgap may be filled. Two-step processes of this type have been analyzed in the small-signal [46] and the steady state regimes [101], but not for the large-signal non-steady-state case in which we are interested.

Fig. 3.6 demonstrates the two-step process. During stress, the Fermi level rises (b), filling the fast surface states, which supply electrons that can tunnel into oxide traps. When the stress is removed (b), the fast surface states emit electrons to the conduction band, providing empty states into which the electrons can tunnel from oxide traps.

It is also possible, in the case of sufficiently high oxide trap concentrations, for tunneling or thermally-activated hopping [133, 75, 74, 134] conduction to occur between oxide traps. We neglect this possibility for now and justify our assumption in Chap. 7 based on the oxide trap densities required to fit the measured data. These values range from 10^{16}–10^{18} cm^{-3} eV^{-1}, corresponding to an average distance between oxide traps of 100–500Å, too large for tunneling between oxide traps to be significant.

The energy of the oxide trap need not be the same as the energy of the interface state, due to inelastic tunneling or lattice relaxation. In order to simplify the following discussion, however, we shall assume only elastic tunneling transitions between interface states and oxide traps at the same energy.

3.5.1 Tunneling Time Constant

The appropriate densities of states, in this case, are

$$\rho_l(E) = D_{it}(E)A \quad (\text{eV}^{-1}) \tag{3.37}$$

$$\rho_r(E,x) = D_{ot}(E,x) \quad (\text{cm}^{-3}\,\text{eV}^{-1}) \tag{3.38}$$

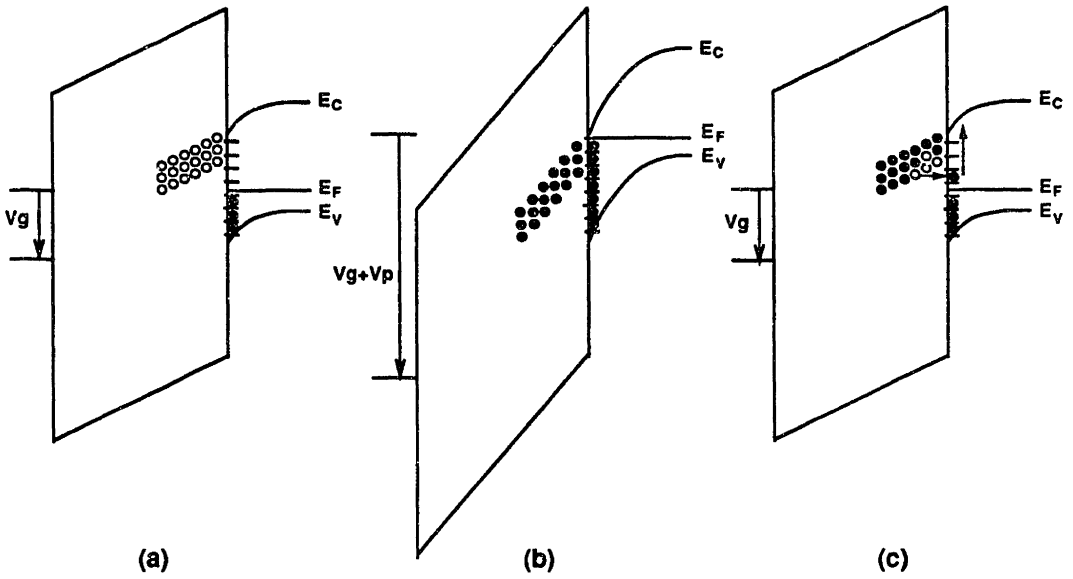


Figure 3.6: Interface trap-assisted tunneling. (a) Before stress, (b) During stress; oxide traps fill by tunneling from interface states, (c) After stress; oxide traps empty by tunneling to empty interface states.

where A is the area of the interface and D_{it} ($\mathrm{cm}^{-2}\ \mathrm{eV}^{-1}$) is the interface state density. In this case, the occupancy functions are

$$f_l = f_{it}(E,t) \tag{3.39}$$

$$f_r = f_{ot}(E,x,t) \tag{3.40}$$

where f_{it} and f_{ot} are the occupancy functions for interface states and oxide traps, respectively, which in general depend on time.

Subsituting Eqs. 3.37–3.40 into Eq. 3.10, the time rate of change of the occupancy function for oxide traps at energy E and depth x into the oxide is,

$$\frac{\partial f_{ot}}{\partial t} + \frac{f_{ot}}{\tau_{ot/it}} = \frac{f_{it}}{\tau_{ot/it}} \tag{3.41}$$

The time constant for tunneling between oxide traps and interface states is obtained from Eq. 3.9 using Bardeen's effective matrix element [Eq. 3.16]. Since electrons trapped in interface states do not possess momentum along the interface plane, $\vec{k}_{||} = 0$, and the summation over $\vec{k}_{||}$ vanishes. Thus,

$$\frac{1}{\tau_{ot/it}} = \omega_{ot/it} = \frac{2\pi}{\hbar}|M_{lr}|^2 D_{it} A \tag{3.42}$$

The time constant $\tau_{ot/it}$ is evaluated from Eq. 3.42 in Appendix B. The result is

$$\tau_{ot/it}(E,x) = \tau_o(E,x)\exp\lambda(x) \tag{3.43}$$

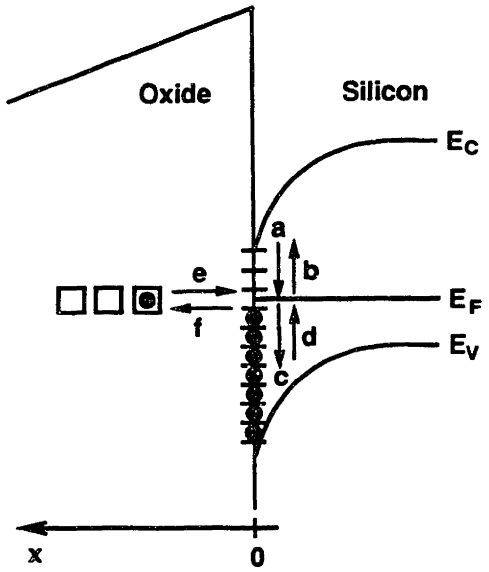


Figure 3.7: Two-step tunneling between oxide traps and interface states showing possible electron transitions.

We defer the explicit evaluation of the WKB integral $\lambda(x) = 2\int_0^x K(x')dx'$ until the following section, since two-band effects on the oxide dispersion relation may be important in this case. The prefactor $\tau_o(E,x)$ again depends on the model used for the oxide trap potential. In the case of a one-dimensional delta function trap potential

$$\frac{1}{\tau_o(E,x)} = \frac{4\pi\hbar^3\sigma_n D_{it}}{M^{*2}} \frac{\kappa(0)K(0)^2K(x)^2}{\kappa(0)+K(0)} \tag{3.44}$$

where M^* is the effective mass in the oxide and κ and K are defined in Eqs B.104 and B.105. For a three-dimensional spherical delta function potential well,

$$\frac{1}{\tau_o(E,x)} = \frac{8\pi^2\hbar^3 D_{it}(E)}{M^{*2}} \left(\frac{\kappa K^2}{\kappa+K}\right) \tag{3.45}$$

Equations 3.44 and 3.45 are equal, in the case of a rectangular barrier, if the capture cross section has the value

$$\sigma_n = \frac{2\pi}{K^2}$$

3.5.2 Two-step Process

Since the interface states are described by SRH kinetics, it is convenient at this point to recast Eq. 3.41 into a kinetic rate equation of the same form. In this way,

the equations can be formulated as a generalized set of SRH equations which take into account tunneling transitions from interface states to oxide traps. Referring to Fig. 3.7, the time rate of change of the density of occupied oxide traps n_{ot} (cm^{-3} eV^{-1}) at energy E between x and $x + \Delta x$ due to transitions to and from interface states at the same energy is

$$\frac{\partial n_{ot}}{\partial t} = r_f(E,x) - r_e(E,x) \tag{3.46}$$

where $r_e(E,x)$ is the rate ($\mathrm{cm}^{-3}\mathrm{eV}^{-1}\mathrm{s}^{-1}$) at which electrons tunnel from oxide traps into interface states, given by

$$r_e(E,x) = T_{x\to 0}(E,x) n_{ot}(E,x) p_{it}(E)$$

where $p_{it} = (1 - f_{it})D_{it}$ is the density of unfilled interface states. The rate of the reverse process is

$$r_f(E,x) = T_{0\to x}(E,x) n_{it}(E) p_{ot}(E,x)$$

where $n_{ot}(E,x) - f_{ot}D_{ot}$ and $p_{ot}(E,x) = (1 - f_{ot})D_{ot}$ are, respectively, the densities of occupied and unoccupied oxide traps (cm^{-3} eV^{-1}) at E and x and the constants of proportionality, $T_{0\to x}$ and $T_{x\to 0}$ have units of $\mathrm{cm}^2\mathrm{eVs}^{-1}$. Thus,

$$\begin{aligned}\frac{\partial n_{ot}}{\partial t} &= T_{0\to x} p_{ot} n_{it} - T_{x\to 0} n_{ot} p_{it} \\ &= D_{ot} D_{it} [T_{0\to x} f_{it}(1 - f_{ot}) - T_{x\to 0} f_{ot}(1 - f_{it})]\end{aligned} \tag{3.47}$$

In equilibrium, we have detailed balance; $\frac{\partial n_{ot}}{\partial t} = 0$ and $f_{it} = f_{ot}$. Equation 3.47 then implies that $T_{0\to x} = T_{x\to 0} = T(E,x)$. Thus,

$$\frac{\partial f_{ot}}{\partial t} = T(E,x) D_{it}(E) [f_{it}(E) - f_{ot}(E)] \tag{3.48}$$

Equation 3.48 is identical to the previous result, Eq. 3.41, where we identify,

$$\frac{1}{\tau_{ot/it}} = \omega_{ot/it}(E,x) = T(E,x) D_{it}(E) \tag{3.49}$$

Equation 3.48 can be rewritten in a form that resembles the Shockley-Read-Hall rate equations,

$$\frac{\partial n_{ot}}{\partial t} = T(E,x)[D_{ot}(E,x) n_{it}(E,t) - n_{ot}(E,x,t) D_{it}(E)] \tag{3.50}$$

The thermal process is described by the usual SRH rate equations. The time rate of change of the surface concentrations of electrons and holes are given by

$$\begin{aligned}\frac{\partial n_s}{\partial t} &= r_b - r_a \\ &= e_{ns} n_{it} - c_{ns} n_s p_{it}\end{aligned} \tag{3.51}$$

$$\begin{aligned}\frac{\partial p_s}{\partial t} &= -r_c + r_d \\ &= -c_{ps} p_s n_{it} + e_{ps} p_{it}\end{aligned} \tag{3.52}$$

Where $c_{ns} = \sigma_{no}\overline{v}$ and $c_{ps} = \sigma_{po}\overline{v}$, with $\overline{v}$ the thermal velocity and σ_{no} and σ_{po} the electron and hole capture cross sections, respectively. Applying detailed balance, $\frac{\partial n_s}{\partial t} = 0$ and $\frac{\partial p_s}{\partial t} = 0$ so that the emission coefficients can be written $e_{ns} = c_{ns}n_1$ and $e_{ps} = c_{ps}p_1$, where

$$n_1(E) = n_{s0}e^{(E-E_F)/kT} = N_c \exp[(E - E_c)/kT] \quad (3.53)$$

$$p_1(E) = p_{s0}e^{(E_F-E)/kT} = N_v \exp[(E_v - E)/kT] \quad (3.54)$$

where n_{s0} and p_{s0} are the equilibrium electron and hole concentrations and the second equality holds only in the Boltzmann approximation ($E_v + 3kT < E_F < E_c - 3kT$), when

$$n_{s0} = N_c e^{(E_F-E_c)/kT} \quad (3.55)$$

$$p_{s0} = N_v e^{(E_v-E_F)/kT} \quad (3.56)$$

The charge trapped in interface states changes due to thermal transitions of electrons and holes with the silicon conduction and valence bands and also by tunneling to and from oxide traps.

$$\frac{\partial n_{it}}{\partial t} = \left.\frac{\partial n_{it}}{\partial t}\right|_{thermal} + \left.\frac{\partial n_{it}}{\partial t}\right|_{tunnel} \quad (3.57)$$

Using the notation defined in Fig. 3.6,

$$\begin{aligned}\left.\frac{\partial n_{it}}{\partial t}\right|_{thermal} &= r_a - r_b - r_c + r_d \\ &= c_{ns}n_s p_{it}(E,t) - e_{ns}n_{it}(E,t) - c_{ps}p_s n_{it} + e_{ps}p_{it}(E,t) \\ &= -n_{it}(E,t)[c_{ns}(n_s + n_1) + c_{ps}(p_s + p_1)] + D_{it}(E)[c_{ns}n_s + c_{ps}p_1]\end{aligned} \quad (3.58)$$

Since we are assuming only elastic tunneling, the second term in Eq. 3.57 results from tunneling transitions between interface states and oxide traps at the same energy. The oxide traps, however, may be located spatially anywhere within the insulator. Due to the microscopic graininess of the trap distribution, oxide traps will in general not line up with interface states in a direction perpendicular to the interface and tunneling transitions will take place along three spatial dimensions. For analytical simplicity, we assume a macroscopically averaged oxide trap distribution in space which is a continuous function of energy and one spatial dimension perpendicular to the interface. We consider the one-dimensional problem of electrons tunneling into the oxide along the direction x. The total rate of change of the interface trapped charge density is obtained by integrating Eq. 3.50 over all space,

$$\begin{aligned}\left.\frac{\partial n_{it}}{\partial t}\right|_{tunnel} &= -\int_0^{t_{ox}} \frac{\partial n_{ot}}{\partial t} dx \\ &= \int_0^{t_{ox}} T(E,x)[n_{ot}(E,x,t)D_{it}(E) - D_{ot}(E,x)n_{it}(E,t)]dx\end{aligned} \quad (3.59)$$

where t_{ox} is the thickness of the insulator.

The rate equations 3.50, 3.51, 3.52, 3.57, 3.58 and 3.59 can be summarized as

$$\frac{\partial n_s}{\partial t} = c_{ns}(n_s + n_1)n_{it} - c_{ns}D_{it}n_s \tag{3.60}$$

$$\frac{\partial p_s}{\partial t} = -c_{ps}(p_s + p_1)n_{it} + c_{ps}p_1 D_{it} \tag{3.61}$$

$$\begin{aligned}\frac{\partial n_{it}(E,t)}{\partial t} &= -n_{it}(E,t)[c_{ns}(n_s + n_1) + c_{ps}(p_s + p_1)] + D_{it}(E)[c_{ns}n_s + c_{ps}p_1] \\ &\quad + \int_0^{t_{ox}} T(E,x)[n_{ot}(E,x,t)D_{it}(E) - D_{ot}(E,x)n_{it}(E,t)]dx\end{aligned} \tag{3.62}$$

$$\frac{\partial n_{ot}(E,x,t)}{\partial t} = -T(E,x)n_{ot}(E,x,t)D_{it}(E) + T(E,x)n_{it}(E,t)D_{ot}(E,x) \tag{3.63}$$

A complete description of the trapping dynamics under arbitrary bias conditions requires the solution of this system of four coupled nonlinear differential equations. While a numerical solution to the full system 3.60-3.63 is possible, it is not practical for several reasons. In order to obtain the total response, whether a threshold voltage shift as in Chapter 4 or a recovery voltage as in Chapter 5, the solution for $n_{ot}(E,x,t)$ obtained from these equations must be integrated over all energies and depths. For an arbitrary energetic and spatial distribution of oxide traps, this integration must be performed numerically, at each time point. If, in addition, the integrand of this two-dimensional integration must be obtained from a numerical solution to a system of four differential equations, the resulting computational burden is formidable. The situation is complicated by the fact that the equations are nonlinear. The numerical approach also completely obscures any qualitative understanding of the trapping dynamics. We will therefore appeal to several simplifying assumptions in order to make the equations more tractable.

3.5.3 Assumptions

The first assumption is that the system is initially in equilibrium so that electrons, holes, interface states and oxide traps share a common Fermi level E_F. Since oxide traps may have extremely large time constants, this assumption requires some further justification. The traps of interest are those which change occupancy on time scales comparable to time scales t_m of the measurement. Since oxide traps fill or empty, during a time t_m, to a maximum depth on the order of $x_m \approx x_o \ln(t_m/\tau_o)$, from Eq. 2.2, it is only traps with $x < x_m$ whose occupancies change significantly during the measurement sequence. As long as sufficient time is allowed between pulses so that oxide traps at depths $x < x_m$ can equilibrate with electrons, holes and fast surface states at the interface, then all these states can all be described by the same quasi-Fermi level. This condition is easily monitored in practice as the time required for the relaxation, whether a threshold voltage in a MOSFET or a recovery voltage

in a capacitor, to cease changing with time. Traps at depths greater than x_m will not, in general, reach equilibrium during the measurement and hence, it is not valid to extend the Fermi level into the oxide for $x > x_m$. Since their occupancy changes negligibly during the course of the measurement, these spatially deep traps contribute to a constant flatband voltage shift but not to time-dependent threshold voltage shifts. Thus, if the time between pulses is t_m, it is valid to assume that, immediately prior to the next pulse, oxide traps at depths $x < x_m$ are in equilibrium with the interface with an occupancy determined by the equilibrium Fermi level at the interface at the pre-stress bias conditions.

Immediately after application of a step in voltage across the structure, the equilibrium Fermi levels will split, with electrons, holes, interface states and oxide traps each described by separate quasi-Fermi levels; $F_n(t)$, $F_p(t)$, $F_{it}(t)$, $F_{ot}(x,t)$, respectively. As described above, traps at depths $x < x_m$ will will return to equilibrium in time for the next pulse, at which time the quasi-Fermi levels will merge again. Traps cannot respond instantaneously however, and for some time following stress the system will be in a non-steady-state condition in which the quasi-Fermi levels do not line up and change with time. We are interested in solving for the non-steady-state occupancy function of the interface states and oxide traps, defined by $n_{it}(E,t) = f_{it}(E,t)D_{it}(E)$ and $n_{ot}(E,x,t) = f_{ot}(E,x,t)D_{ot}(E,x)$, where

$$f_{it}(E,t) = \frac{1}{1+\exp[(E-F_{it}(x,t))/kT]} \tag{3.64}$$

$$f_{ot}(E,x,t) = \frac{1}{1+\exp[(E-F_{ot}(x,t))/kT]} \tag{3.65}$$

The second assumption we make is that, following a change in gate voltage, electrons and holes at the interface equilibrate rapidly in comparison to the equilibration time for interface states and oxide traps. The electron and hole concentrations immediately following a change in gate voltage can then be calculated assuming equilibrium, $F_n = F_p = E_F$, with n_{it} and n_{ot} at their pre-stress values. As charge is removed from the inversion layer by trapping in interface states and oxide traps, the electron and hole concentrations will readjust with time. If the electron and hole equilibration times are very much less than the equilibration time for traps, then a quasi-static solution to Eqs. 3.60–3.63 can be obtained by by partioning time into increments of duration Δt and solving interatively. At the beginning of each Δt increment, n_s and p_s are calculated from Eqs. E.11 and E.12 with a Fermi level obtained using the current values of n_{it} and n_{ot}. These values of the carrier concentrations can then be used in Eqs. 3.62 and 3.63 to compute $\partial n_{it}/\partial t$ and $\partial n_{ot}/\partial t$ and hence to project n_{it} and n_{ot} to the beginning of the next time interval.

A further simplification can be made in strong inversion, which is of greatest interest for circuit applications, if the sum of the interface and oxide trap concentrations is much less than the total inversion layer carrier concentration. Since charge trapping

then only weakly perturbs the inversion charge concentration, the number of inversion layer charge carriers changes negligibly with time and is well approximated by its value immediately following stress. This assumption will be made in the following development.

In order for the above approximation to be valid, we must first justify the quasi-equilibrium assumption, in which n_s and p_s reach equilibrium much more rapidly than n_{it} and n_{ot}. The capacitors used in our experiments have degenerately doped plates so that carriers can be supplied rapidly, within the dielectric relaxation time, in response to a change in gate voltage. Thus, the free carrier concentration equilibrates rapidly relative to time scales of interest. In MOSFETs, the source drain regions can rapidly supply minority carriers in inversion, and the substrate can supply majority carriers in accumulation. The device is therefore never driven into deep depletion, and quasi-equilibrium conditions will hold. The only time this approximation may fail is when the device is driven into depletion or weak inversion, when carriers must exit the channel by diffusion. Nevertheless, the trapping time constants which are measured in our experiments (microsecond-millisecond range) are sufficiently long for quasi-equilibrium to be a very good approximation.

With these assumptions, the system is reduced from the original four equations, Eqs. 3.60–3.63, to two, Eqs. 3.62 and 3.63. In the following sections, we discuss several methods for solving these equations.

3.5.4 Numerical Solution

With n_s and p_s assumed constant, Eqs. 3.62 and 3.63 are linear, so that a numerical solution is efficient. The equations can be written in discrete form by subdividing the oxide into N slices of width Δx,

$$\begin{aligned}\frac{dn_{it}}{dt} &= -n_{it}\left[c_{ns}(n_s+n_1)+c_{ps}(p_s+p_1)+\sum_{i=1}^{N}T(x_i)D_{ot}(x_i)\Delta x\right] \\ &\quad +D_{it}\sum_{i=1}^{N}T(x_i)n_{ot}(x_i)\Delta x + D_{it}\left[c_{ns}n_s+c_{ps}p_1\right] \qquad (3.66)\end{aligned}$$

$$\frac{d}{dt}\left[n_{ot}(x_i)\Delta x\right] = T(x_i)\left[D_{ot}(x_i)\Delta x\right]n_{it} - T(x_i)D_{it}\left[n_{ot}(x_i)\Delta x\right] \qquad (3.67)$$

for $i = 1,\ldots,N$. Equations 3.66 and 3.67 can be written more concisely as

$$\dot{y} = Cy + d$$

with

$$\begin{aligned}y &= [n_{it}, n_{ot}(x_1)\Delta x,\ldots,n_{ot}(x_N)\Delta x]^T \\ d &= [D_{it}(c_{ns}n_s+c_{ps}p_1),0,0,\ldots,0]^T\end{aligned}$$

$$C = \begin{bmatrix} -a_1 & T(x_1)D_{it} & \cdots & T(x_N)D_{it} \\ T(x_1)D_{ot}(x_1)\Delta x & -T(x_1)D_{it} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ T(x_N)D_{ot}(x_N)\Delta x & 0 & \cdots & -T(x_N)D_{it} \end{bmatrix}$$

$$a_1 = c_{ns}(n_s + n_1) + c_{ps}(p_s + p_1) + \sum_{i=1}^{N} T(x_i)D_{ot}(x_i)\Delta x$$

This is a stiff set of equations with widely spaced eigenvalues which is most easily solved by the forward Euler method,

$$\begin{aligned} y_{n+1} &= y_n + \Delta t \dot{y}_{n+1} \\ &= y_n + \Delta t(Cy_{n+1} + d) \qquad (3.68) \\ y_{n+1} &= (1 - C\Delta t)^{-1}(y_n + d\Delta t) \qquad (3.69) \end{aligned}$$

Equation 3.69 can be solved numerically for the vector of oxide and interface trap occupancies at time $(n+1)\Delta t$ given their values at time $n\Delta t$.

The system 3.66 and 3.67 has N eigenvalues and hence N time constants, which approach a continuous distribution as $N \to \infty$. Thus an exact solution involves a continuum of time constants and is not useful. In order to gain some insight into the effect of the interface state dynamics on the charging and discharging of oxide traps we consider several approximations.

3.5.5 Approximate Solutions

Approximation 1: Decoupled System

Under most normal bias conditions the thermal (SRH) time constant is very much smaller than the tunneling time constant. Under the assumption that the two time constants are widely separated, Eqs. 3.62 and 3.63 can be decoupled. This amounts to neglecting tunneling in the n_{it} calculation and assuming that the interface states change occupancy instantaneously in the calculation of n_{ot}. For example, after the application of a positive stress voltage to an n-channel MOSFET, the Fermi level rises and fast surface states at the interface fill "instantaneously" with electrons before the oxide traps have time to respond. These filled interface states then provide a constant supply of electrons which can tunnel into oxide traps. When the stress is removed, electrons are just as rapidly swept out of the interface states to the silicon bands, providing a constant supply of empty states into which charge can tunnel from the filled oxide traps.

Neglecting tunneling in Eq. 3.62, the non-steady-state occupancy function for the interface states is

$$f_{it}(E,t) = f(E) + [f_o(E) - f(E)]e^{-t/\tau_{it}} \qquad (3.70)$$

where the time constant is

$$\tau_{it} = \frac{1}{c_{ns}(n_s + n_1) + c_{ps}(p_s + p_1)} \tag{3.71}$$

and f_o and f are the steady-state occupancy functions before and after stress, respectively, given by

$$f(E) = \frac{\sigma_{no} n_s + \sigma_{po} p_1}{\sigma_{no}(n_s + n_1) + \sigma_{po}(p_s + p_1)} \tag{3.72}$$

$$f_o(E) = \frac{\sigma_{no} n_{s0} + \sigma_{po} p_{10}}{\sigma_{no}(n_{s0} + n_{10}) + \sigma_{po}(p_{s0} + p_{10})} \tag{3.73}$$

The steady-state occupancy functions $f_o(E)$ and $f(E)$, which describe the fractional occupation of interface and oxide traps before stress ($t = 0^-$) and after stress ($t \rightarrow \infty$), respectively, are in general, not the same as the equilibrium Fermi functions. From Eq. 3.72, $(1-f)/f = \exp[(E-F)/kT]$, so that from Eq. 3.64 the steady-state interface state quasi-Fermi level is

$$F_{it} = E_{it} + kT \ln\left(\frac{\sigma_{no} n_s + \sigma_{po} p_1}{\sigma_{no} n_1 + \sigma_{po} p_s}\right) \tag{3.74}$$

In equilibrium, substituting Eqs. 3.53-3.56 into Eq. 3.74, we find $F_{it} = E_F$. Thus, under the assumption that the carrier concentrations can be calculated assuming equilibrium conditions, the steady-state occupancy functions are identical to the equilibrium Fermi functions

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]} \tag{3.75}$$

$$f_o(E) = \frac{1}{1 + \exp[(E - E_{F0})/kT} \tag{3.76}$$

with E_{F0} and E_F the Fermi levels before and during stress.

Since $\tau_{it} \ll \tau_{ot/it}$, n_{it} is approximately constant at its final value $n_{it}(t \rightarrow \infty) = fD_{it}$ during the filling of the oxide traps. Sufficient time is allowed between pulses so that the initial value of n_{ot} is determined by the steady-state value of n_{it} prior to stress, $n_{it}(E, 0) = f_o D_{it}$, so that, from Eq. 3.63 the initial condition for the oxide traps is $n_{ot}(E, x, 0) = f_o D_{ot}(E, x)$. This implies that oxide traps are allowed to reach equilibrium with the interface states between pulses, an assumption which has already been justified.

During and after application of a pulse, the traps are in a nonequilibrium, non-steady-state condition, and the quasi-Fermi level $F_{ot}(x, t)$ differs from its equilibrium value and varies with time. Solving Eq. 3.63 with this initial condition, the non-steady state oxide trap occupancy function is

$$f_{ot}(E, x, t) = f(E) + [f_0(E) - f(E)]e^{-t/\tau_{ot/it}} \tag{3.77}$$

where

$$\tau_{ot/it}(E,x) = \frac{1}{\omega_{ot/it}(E,x)} = \frac{1}{T(E,x)D_{it}(E)} \tag{3.78}$$

Note that, in contrast to the time constant for the direct mechanism in Sec. 3.4.1, which was inversely proportional to the density of conduction band states, the tunneling time constant for the two-step process is inversely proportional to the interface state density $D_{it}(E)$.

The assumption of widely separated time constants is normally valid at room temperature and in strong inversion. As shown in Section 2.1.3, however, there may be some overlap in the time constants at low temperatures or when the Fermi level is near midgap when τ_{it} is large. Clearly, this approximation will fail to model any temperature dependence of n_{ot} in the two-step process, because the tunneling is independent of temperature and the dynamics of the thermal SRH process have been eliminated.

Approximation 2: Partially Coupled System

An improved approximation can be obtained by taking account of the dynamics of the interface states, to first order, in the calculation of n_{ot}. We continue to neglect tunneling in the calculation of n_{it}, under the assumption that, of the two processes which act in parallel to change the occupancy of the interface states, tunneling is normally much slower than the thermal process. However, in the calculation of $n_{ot}(E,x,t)$ we use the non-steady-state expression for $f_{it}(E,t)$, Eq. 3.70, rather than its steady-state final value Eq. 3.72. Thus, substituting Eq. 3.70 into the equation for the oxide traps,

$$\begin{aligned} \frac{\partial f_{ot}}{\partial t} + \frac{f_{ot}}{\tau_{ot/it}} &= TD_{it}f_{it} \\ &= \frac{1}{\tau_{ot/it}}\left[f + (f_o - f)e^{-t/\tau_{it}}\right] \end{aligned} \tag{3.79}$$

Assuming $f_{ot}(0) = f_o$, Eq. 3.79 can be solved to give the occupancy function for the oxide traps

$$f_{ot}(E,x,t) = f(E) + [f_o(E) - f(E)]\left[\left(\frac{\tau_{ot/it}}{\tau_{ot/it} - \tau_{it}}\right)e^{-t/\tau_{ot/it}} - \left(\frac{\tau_{it}}{\tau_{ot/it} - \tau_{it}}\right)e^{-t/\tau_{it}}\right] \tag{3.80}$$

The response of the oxide traps is a weighted sum of the "uncoupled" responses, characterized by the two time constants τ_{it} and $\tau_{ot/it}$. If $\tau_{ot/it} \gg \tau_{it}$, Eq. 3.80 reduces to the previous result, Eq. 3.77. If $\tau_{it} \gg \tau_{ot/it}$, the the interface states are rate limiting and the response is controlled by τ_{it},

$$f_{ot} = f + (f_o - f)e^{-t/\tau_{it}}$$

This may be important at low temperatures or in depletion when τ_{it} becomes large (see discussion in Sec. 2.1.3). Note that Eq. 3.80 is also well behaved when $\tau_{it} = \tau_{ot/it}$. With the exception of the case in which the two time constants are very close in value, Eq. 3.80 looks like a first order response governed by the larger of the two time constants τ_{it} and $\tau_{ot/it}$. In either case, after a sufficiently long time the oxide traps fill to a level $n_{ot}(t \rightarrow \infty) = fD_{ot}$. Note that Eq. 3.80 is temperature dependent through the dependence of τ_{it} on the carrier concentrations in Eq. 3.71.

Approximation 3: Delta Function Spatial Distribution of Traps

The two-step tunneling problem can be solved exactly in the case of a distribution of traps concentrated at one point in space with density n_{ot} (cm^{-2} eV^{-1}). In contrast to the exact problem which has an infinite number of time constants, this system has only two and is easily solved in closed form. As shown in Appendix C, the non-steady-state occupancy functions for interface states and oxide traps are

$$\begin{aligned} f_{it}(E,t) &= f(E) + [f_o(E) - f(E)]\left[\left(\frac{\tau_1}{\tau_1 - \tau_2}\right)e^{-t/\tau_1} - \left(\frac{\tau_2}{\tau_1 - \tau_2}\right)e^{-t/\tau_2}\right] \\ f_{ot}(E,x,t) &= f(E) + [f_o(E) - f(E)]\left[\left(\frac{\tau_1 - \tau_{ot/it}}{\tau_1 - \tau_2}\right)e^{-t/\tau_1} - \left(\frac{\tau_2 - \tau_{ot/it}}{\tau_1 - \tau_2}\right)e^{-t/\tau_2}\right] \end{aligned} \tag{3.81}$$

where τ_1 and τ_2 are given by

$$\tau_{1,2} = \frac{1}{2}(\tau_{it} + \tau_{ot/it} + \tau_{it}\tau_{ot/it}Tn_{ot})\left\{1 \pm \left[1 - \frac{4\tau_{it}\tau_{ot/it}}{(\tau_{it} + \tau_{ot/it} + \tau_{it}\tau_{ot/it}Tn_{ot})^2}\right]^{1/2}\right\} \tag{3.82}$$

with $f(E)$ and $f_o(E)$ given by Eqs. 3.72 and 3.73.

Approximation 4: First Order Solution

The final approximation we consider is the simplest of all and is based on the observation that the tunneling and thermal transitions in the two-step process occur in series. It seems reasonable to write the total time constant for the two-step mechanism as

$$\tau = \tau_{ot/it} + \tau_{it} \tag{3.83}$$

and to model the process by a first-order rate equation,

$$f_{ot}(E,x,t) = f(E) + [f_o(E) - f(E)]e^{-t/\tau} \tag{3.84}$$

This equation is not rigorously correct since the actual process is characterized by a second order differential equation with two time constants. While Eqs. 3.83 and 3.84 have the correct aysmptotic behavior when $\tau_{ot/it}$ and τ_{it} are widely separated, they may deviate from the exact solution when the time constants are comparable in magnitude.

In the next section we investigate the relative accuracy of these approximations.

Comparison of Exact and Approximate Solutions

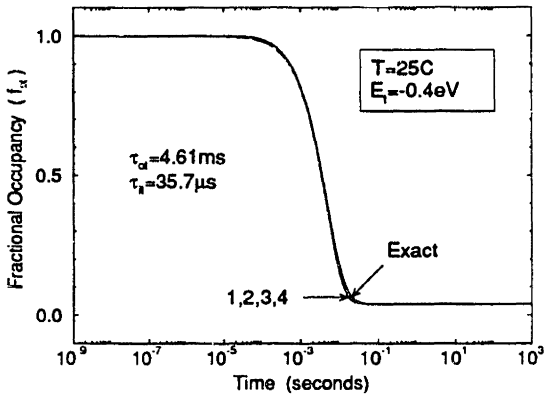
This section presents simulation results which compare the approximate solutions derived above with the exact solution obtained numerically from Eq. 3.69. The examples which follow are for the case of an NMOS transistor with a 250$\AA$ gate oxide and an initial gate voltage of 5V. The gate voltage is then stepped to 1V ($V_p = -4$V) and the fractional occupancy of traps at various energies and depths is observed as a function of time. This bias condition was chosen because it drives the electron and hole concentrations at the interface to near-intrinsic levels which maximize the value of the thermal time constant τ_{it} [Eq. 3.71].

Approximation 1 [Eq. 3.77] completely neglects the thermal process, assuming that tunneling is always the rate–limiting step. This assumption breaks down as the thermal time constant increases relative to the tunneling time constant, i.e. for decreasing temperature, for energies approaching midgap and for oxide traps very close to the interface. These effects are illustrated in the examples which follow.

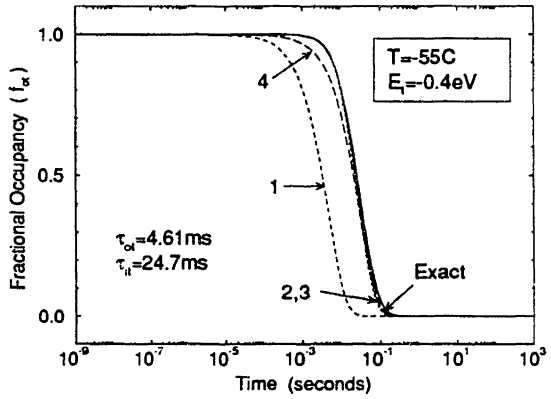
Fig. 3.8 shows emission transients for an oxide trap at a depth of 10$\AA$. For an oxide trap 0.4eV below the bottom of the conduction band, the approximate solutions all show excellent agreement with the exact calculation at 25°C [Fig. 3.8(a)] when $\tau_{ot/it} \gg \tau_{it}$. At −55°C [Fig. 3.8(b)], $\tau_{it} > \tau_{ot/it}$ and approximation 1 greatly underestimates the actual emission time, which is limited by τ_{it}. Approximation 4 makes a good first-order estimate of the effective time constant for the discharge, but the dynamics start to deviate from the exact solution when the time constants are comparable. Approximations 2 and 3 show good agreement with the exact solution even at the lower temperature. Figure 3.8(c) shows the same conditions as in Fig. 3.8(b) except that the interface state energy is 0.1eV below the conduction band. At this energy the SRH time constant is sufficiently short that all the approximations, even the decoupled solution, are satisfactory, even at −55°C.

Figure 3.9 shows emission transients at −55°C for an oxide trap at energy 0.4eV below the conduction band and various distances from the interface. At a depth of 15$\AA$ as shown in Fig. 3.9(a), $\tau_{ot/it} \gg \tau_{it}$ and all of the approximated solutions show good agreement with the exact calculation. At a depth of 10$\AA$, however, $\tau_{ot/it} < \tau_{it}$ and the thermal process becomes rate-limiting. Approximation 1 again underestimates the emission time. In approximation 4, the traps discharge on the correct time scale but there are deviations in the functional form of the decay. As the oxide trap moves closer to the interface as in Fig. 3.9(c), the decoupled approximation greatly underestimates the actual emission time while the exact emission transient, which is thermally limited, does not change.

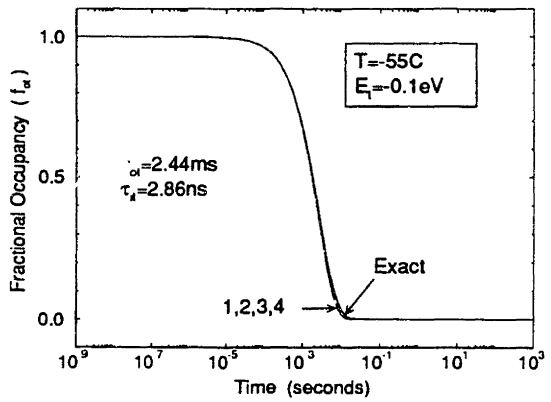
Approximation 2 [Eqs. 3.80] shows excellent agreement with the exact solution over all energies, temperatures and depths for the trap densities used in the above examples. This approximation begins to deviate from the exact solution, however, as the oxide trap density D_{ot} increases relative to the interface state density D_{it}. Figures 3.10(a)–(c) show emission transients for an oxide trap at $x = 5\AA$, $E_t =$



(a)

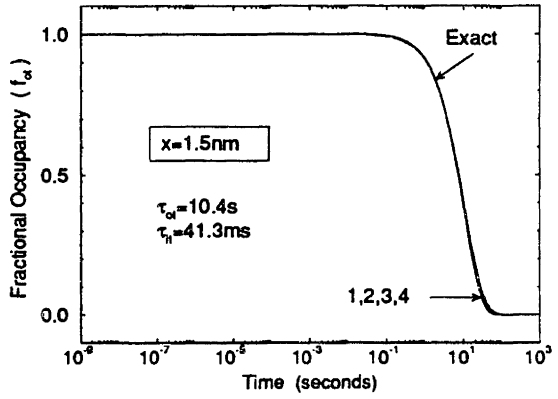


(b)

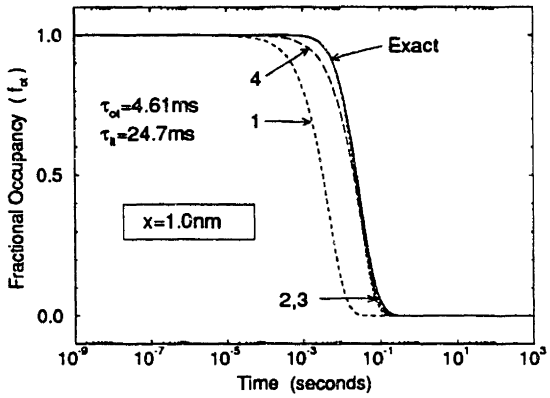


(c)

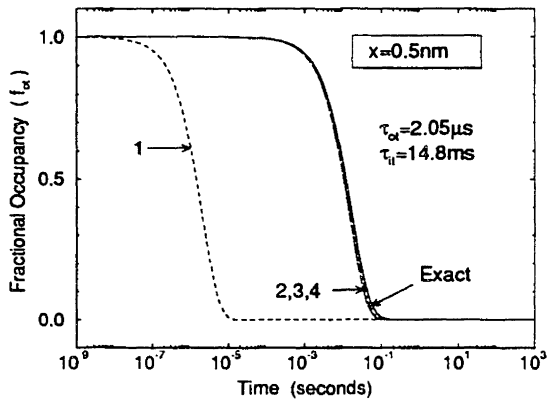
Figure 3.8: Simulations showing oxide trap fractional occupancy as a function of time for $V_G(initial) = 5\text{V}$, $V_G(final) = 1V$, $D_{it} = 1.0 \times 10^{10}\text{cm}^{-2}\ \text{eV}^{-1}$, $D_{ot} = 1.0 \times 10^{17}\text{cm}^{-3}\ \text{eV}^{-1}$, $x = 10\mathring{A}$ and (a) $E_t = -0.4\text{eV}$, $T = 25°\text{C}$ (b) $E_t = -0.4\text{eV}$, $T = -55°\text{C}$ (c) $E_t = -0.1\text{eV}$, $T = -55°\text{C}$.



(a)

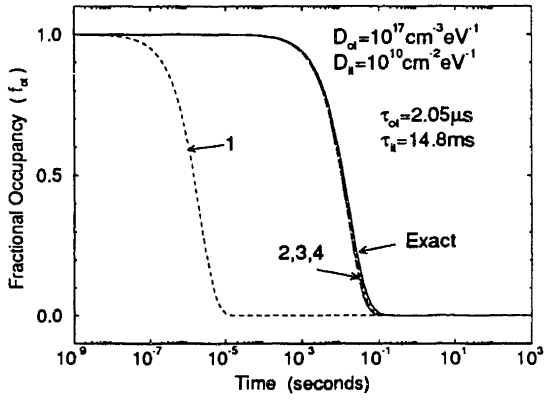


(b)

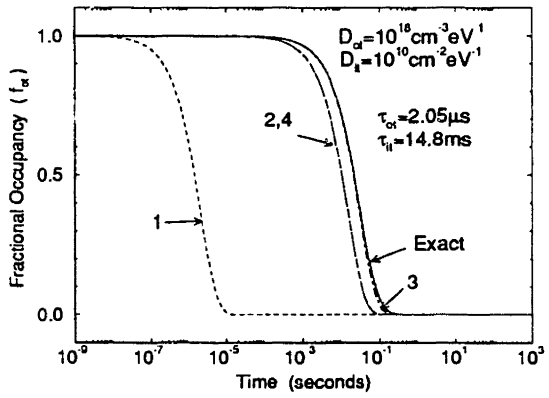


(c)

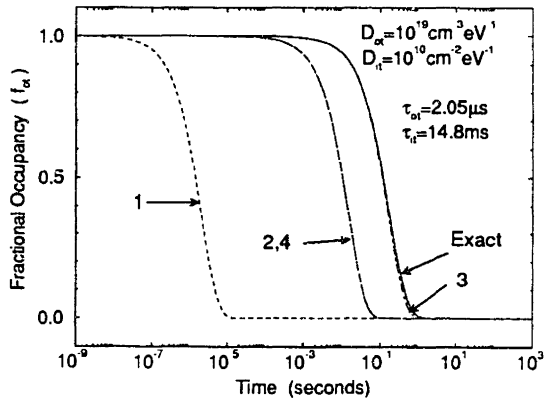
Figure 3.9: Simulations showing oxide trap fractional occupancy as a function of time for $V_G(initial) = 5\text{V}$, $V_G(final) = 1V$, $D_{it} = 1.0 \times 10^{10}\text{cm}^{-2}\ \text{eV}^{-1}$, $D_{ot} = 1.0 \times 10^{17}\text{cm}^{-3}\ \text{eV}^{-1}$, $E_t = -0.4\text{eV}$, $T = -55°\text{C}$ and (a) $x = 15\mathring{A}$ (b) $x = 10\mathring{A}$ (c) $x = 5\mathring{A}$.



(a)



(b)



(c)

Figure 3.10: Simulations showing oxide trap fractional occupancy as a function of time for $V_G(initial) = 5\text{V}$, $V_G(final) = 1V$, $D_{it} = 1.0 \times 10^{10}\text{cm}^{-2}\ \text{eV}^{-1}$, $E_t = -0.4\text{eV}$, $x = 5\mathring{A}$, $T = -55°\text{C}$ and (a) $D_{ot} = 1.0\times10^{17}\text{cm}^{-3}\ \text{eV}^{-1}$ (b) $D_{ot} = 1.0\times10^{18}\text{cm}^{-3}\ \text{eV}^{-1}$ (c) $D_{ot} = 1.0 \times 10^{19}\text{cm}^{-3}\ \text{eV}^{-1}$.

-0.4eV, $T = -55°$C with $D_{it} = 1.0 \times 10^{10}$ cm^{-2} eV^{-1} and $D_{ot} = 1.0 \times 10^{17}$, 1.0×10^{18} and 1.0×10^{19} cm^{-3} eV^{-1}, respectively. Approximation 2 fails at higher oxide trap densities because it neglects the distributed nature of the problem. It is based on an analysis of traps at a single depth in the oxide and assumes that the interface state occupancy is determined only by the thermal process and traps in this one slice of the oxide. In the exact distributed problem, the occupancy of an interface state is also affected by transitions between it and a continuum of traps at various depths in the oxide. Since the availability of empty interface states drives the emission from oxide traps, with increasing D_{ot} relative to D_{it} the actual emission rate is longer than predicted by Eq. 3.80 because of the "crowding" effect of many oxide traps trying to empty through a limited number of interface states.

Approximation 3 [Eq. 3.81] continues to show good agreement with the exact calculation in all cases.

3.6 Two-Band Model

In order to evaluate the values of the imaginary wavevectors K and κ in Eqs. 3.18, 3.43, 3.44 and 3.45 the dispersion relations in the oxide and semiconductor bandgaps are required. In the derivation of the time constant for direct tunneling [Eq. 3.18], we assumed parabolic energy bands so that the attenuation constant in the insulator was given by

$$K^2(x) = K_c^2(x) = \frac{2M_c^*}{\hbar^2}[U_c(x) - E] \tag{3.85}$$

where $U_c(x)$ is the energy of the insulator conduction band minimum and M_c^* is the effective mass of an electron in the conduction band of the insulator, assumed isotropic. This expression results from a Taylor series expansion of the energy to second order in K_c around the insulator conduction band minimum and is therefore strictly valid only for small values of K_c. Since tunneling occurs well within the insulator bandgap, we might inquire as to the validity of the relation 3.85, with a constant value of the effective mass, at energies so far from the band extrema. The same considerations apply to κ in the two-step process, since tunneling from interface states may occur well within the semiconductor bandgap.

Equation 3.85 predicts that, as the energy of the tunneling electron decreases, K and hence the tunneling time constant Eq. 3.43, increase. We expect, however, that for sufficiently low energies, e.g. in the silicon valence band, that the tunneling time constant should begin to decrease again, since holes can tunnel. That is, we expect tunneling to be controlled less by the barrier height between the electron energy and the oxide conduction band, and more by the barrier height to the oxide valence band. Thus, an accurate description requires a two-band model for the attenuation constant K.

In our earlier treatment of direct band-to-trap tunneling, this issue does not present a serious problem. Since tunneling from the silicon bands occurs within a narrow energy range around the band extrema, K does not vary too much and we expect a relation of the form Eq. 3.85 to be a good approximation for some constant value of the effective mass. Since tunneling from the silicon conduction band occurs closer to the conduction band of the oxide than to the valence band, Eq. 3.85 should be valid. Similarly, tunneling from the silicon valence band is assumed to be controlled entirely by the oxide valence band

$$K_v^2(x) = \frac{2M_v^*}{\hbar^2}[E - U_v(x)] \tag{3.86}$$

where M_v^* is the effective mass of a hole in the insulator and $U_v(x)$ is the energy of the insulator valence band maximum. The correct values of the effective masses to use in Eqs. 3.85 and 3.86 however, may differ from their values near the oxide band edges, especially for tunneling from the silicon valence band, which occurs close to the middle of the oxide bandgap, as shown in Fig. 3.11, where the effects of both bands will be important. A completely rigorous treatment should use a two band model for both attenuation constants. However, based on the above considerations, the approximate one-band models are sufficiently accurate for our purposes.

For tunneling at energies within the silicon bandgap, which fall between these two extremes, we must include the effects of both oxide barriers in order to include the possibility of both electron and hole tunneling. Also, the attenuation constant in the semiconductor bandgap κ, which affects the normalization of the wavefunctions, will be strongly influenced by both semiconductor bands. Thus, a two band model is required for the attenuation constants in both the oxide and semiconductor bandgaps for an accurate description of tunneling from the fast surface states.

The effects of nonparabolic energy bands on tunneling characteristics have been considered by several authors [135, 136, 137, 138, 139, 140]. Experimentally, two-band effects have been shown to influence the tunneling current in AlN [140] and in GaAs Schottky barriers [138]. In fact, these tunneling experiments have been used to determine the energy-momentum relationship in the forbidden region.

3.6.1 Attenuation Constant in the Oxide

While several studies of the dispersion relation in SiO_2 have been conducted [141, 142, 143, 144], detailed information is generally not available. Several approximate analytical and empirical $E(k)$ relationships have been proposed [145, 136] and these show reasonable agreement with experiments [138, 140]. A simple approximation uses the parallel combination of the squares of the attenuation constants of the two bands of interest [146],

$$\frac{1}{K^2} = \frac{1}{K_c^2} + \frac{1}{K_v^2}$$

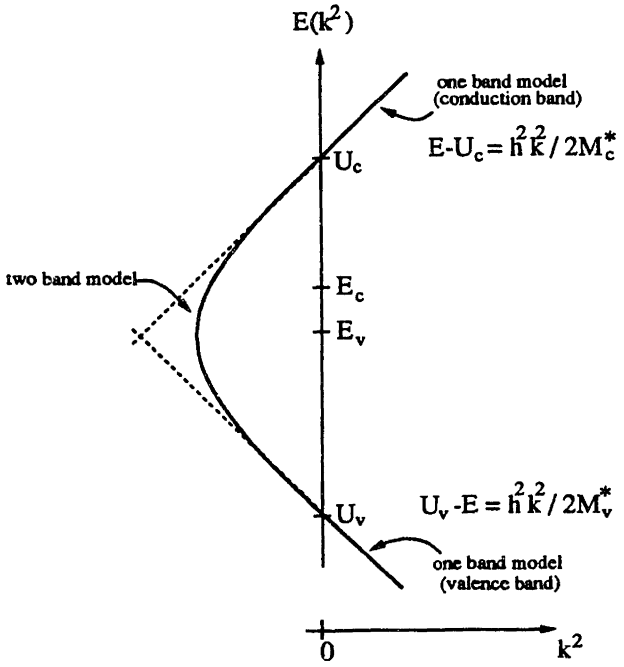


Figure 3.11: Simple two band model for the insulator. $k^2 < 0$ in the energy gap corresponding to decaying exponentials.

$$= \frac{\hbar^2}{2M_c^*(U_c - E)} + \frac{\hbar^2}{2M_v^*(E - U_v)} \tag{3.87}$$

This approximation is shown in Fig. 3.11 and reduces to Eq. 3.85 and Eq. 3.86 in the conduction and valence bands respectively. Solving for K yields

$$K = \left(\frac{2M_c^*}{\hbar^2}\right)^{1/2} [U_c - E]^{1/2} \left[1 - \frac{U_c - E}{E_{Gb}}\right]^{1/2} \left[1 - \frac{U_c - E}{E_{Gb}}\left(1 - \frac{M_c^*}{M_v^*}\right)\right]^{-1/2} \tag{3.88}$$

This expression reduces to the form originally proposed by Franz [145] when the effective masses are equal,

$$\frac{\hbar^2 K^2}{2M^*} = (U_c - E)\left[1 - \frac{U_c - E}{E_{Gb}}\right] \tag{3.89}$$

where $E_{Gb} = U_c - U_v$ is the energy gap of SiO_2, approximately 9.3 eV [147]. In SiO_2, due to the absence of more detailed information, isotropic effective masses for electrons and holes are usually assumed. Good agreement between theoretical and experimental Fowler-Nordheim tunneling results have been found using an isotropic mass ranging from $M_c^* = 0.42m_o$ [143, 142, 141] to $0.5m_o$ [98]. The effective mass M_v^* for holes in SiO_2 ranges from 5 to $10m_o$ [148]. Using the dispersion relation of

Eq. 3.89 and $U_c = \phi_c + q\mathcal{E}x$, the WKB integral in Eq. 3.43 can be evaluated to give [98]

$$\lambda(E) = 2\int_0^x K(E,x)dx \tag{3.90}$$

$$= \frac{(2M_c^*)^{1/2}}{8q\hbar\mathcal{E}} E_{Gb}{}^{3/2}[\sin 4\alpha_o - 4\alpha_o - \sin 4\alpha_1 + 4\alpha_1] \tag{3.91}$$

where

$$\alpha_o = \cos^{-1}\left(\frac{\phi_c - E}{E_{Gb}}\right)^{1/2}$$

$$\alpha_1 = \cos^{-1}\left(\frac{\phi_c - E + q\mathcal{E}x}{E_{Gb}}\right)^{1/2}$$

where $\phi_c = U_c(0) \approx 3.1$eV [149] is the barrier height between the silicon and SiO_2 conduction bands. This result is rather complicated and assumes equal effective masses are in the oxide conduction and valence bands. Simpler results can be obtained by one of several approximations:

Approximation 1

Assuming $U_c - E \ll E_{Gb}$, the integral Eq. 3.90 can be evaluated using Eq. 3.88. Keeping only the first order terms of the Taylor expansions of the square root terms [98],

$$\lambda(x) = \frac{4}{3}\frac{(2M_c^*)^{1/2}}{q\hbar\mathcal{E}}[\xi(\phi_c + q\mathcal{E}x - E) - \xi(\phi_c - E)] \tag{3.92}$$

where

$$\xi(u) = u^{3/2} - \frac{3}{10}\frac{M_c^*}{M_v^*}\frac{u^{3/2}}{E_{Gb}}$$

Approximation 2

Instead of the dispersion relation Eq. 3.88 we can use a simpler approximation, proposed by Esaki [150], in which K at any E (measured with respect to the silicon conduction band) is determined by the smaller of the two barrier heights between the energy of the tunneling electron and the conduction and valence bands of the oxide: For $\phi_c - \frac{M_v^*}{M_c^*+M_v^*}E_{Gb} < E < \phi_v$

$$K(E) = \left[\frac{2M_c^*}{\hbar^2}(\phi_c - E)\right]^{1/2}$$

$$\lambda(x) = \frac{4}{3}\left(\frac{2M_c^*}{\hbar^2}\right)^{1/2}\frac{(\phi_c - E + q\mathcal{E}x)^{3/2} - (\phi_c - E)^{3/2}}{q\mathcal{E}} \tag{3.93}$$

For $\phi_c - E_{Gb} < E < \phi_c - \frac{M_v^*}{M_c^* + M_v^*} E_{Gb}$,

$$\begin{aligned} K(E) &= \left[\frac{2M_v^*}{\hbar^2}(E - \phi_c + E_{Gb})\right]^{1/2} \\ \lambda(x) &= \frac{4}{3}\left(\frac{2M_v^*}{\hbar^2}\right)^{1/2}\left[\frac{(E + E_{Gb} - \phi_c)^{3/2} - (E + E_{Gb} - \phi_c - q\mathcal{E}x)^{3/2}}{q\mathcal{E}}\right] \end{aligned} \tag{3.94}$$

For the effective masses given above, $M_v^* = (10 - 20)M_c^*$, tunneling is essentially controlled by the conduction band over the entire bandgap of the oxide. Thus, we can treat the problem, to very good approximation, from the point of view of electrons tunneling through a barrier governed solely by the oxide conduction band. The effect of the oxide valence band requires only a very minor modification of this picture, which is beyond the accuracy requirements of the current problem.

3.6.2 Attenuation Constant in the Semiconductor

In the semiconductor we assume a Franz dispersion relation of the same form as Eq. 3.87. Taking the zero of energy at the bottom of the semiconductor conduction band, $E_c = 0$,

$$\kappa = \frac{\sqrt{2}}{\hbar}\left[\frac{m_c^* m_v^* E(E + E_G)}{m_c^* E - m_v^*(E + E_G)}\right]^{1/2} \tag{3.95}$$

where $E_G = 1.12\text{eV}$ is the bandgap of silicon. The isotropic effective masses used in Eq. 3.95 are the density of states effective masses in silicon [151],

$$\begin{aligned} m_c^* &= 6^{2/3}(m_l^* m_t^*)^{1/3} \\ m_v^* &= [{m_{hh}^*}^{3/2} + {m_{lh}^*}^{3/2}]^{2/3} \end{aligned}$$

where m_l^* and m_t^* are the longitudinal and transverse effective masses in the conduction band and m_{hh}^* and m_{lh}^* are the effective masses of heavy and light holes, respectively, in the valence band.

The expression 3.95 for κ can be used directly in the equations for the time constants 3.44 and 3.45. Note that $\kappa(E) = 0$ at the conduction band edge $E = 0$ and at the valence band edge $E = -E_G$, a drastically different result from what would have been obtained using a one-band model.

3.7 Summary

In summary, the non-steady-state occupancy function for oxide traps satisfies the equation,

$$\frac{\partial f_{ot}(E,x,t)}{\partial t} + \frac{f_{ot}(E,x,t)}{\tau_{ot/i}} = \frac{f_{ot}(E,x,\infty)}{\tau_{ot/i}} \tag{3.96}$$

with solution

$$f_{ot}(E,x,t) = f_{ot}(E,x,\infty) + [f_{ot}(E,x,0^-) - f_{ot}(E,x,\infty)]R_i(E,x,t) \qquad (3.97)$$

A generalized notation, R_i, has been introduced to describe the time dependent part of the response for transitions between oxide traps and states of type i. This shorthand notation can represent either a first or second-order charging transient and will be used widely in the next chapter. The corresponding tunneling time constant is $\tau_{ot/i}$.

If the initial ($t = 0^-$) and final ($t \to \infty$) states are in equilibrium, then

$$\begin{aligned} f_{ot}(E,x,0^-) &= f_0(E) \\ f_{ot}(E,x,\infty) &= f(E) \end{aligned}$$

where f_0 and f are the equilibrium fermi functions before and during stress, respectively.

The rate equations for oxide traps were solved for band-to-trap and trap-to-trap transitions. The results are summarized below.

Band-to-Trap Transitions ($i \in \{cb, vb\}$):

Tunneling from conduction band states to oxide traps is a simple first-order process:

$$R_{cb}(E,x,t) \equiv e^{-t/\tau_{ot/cb}(E,x)}$$

where, for direct tunneling transitions between a conduction band electron and an oxide trap at the same energy,

$$\tau_{ot/cb} = \tau_o e^{\lambda(x)} \qquad (3.98)$$

$\lambda(x) = 2\int K(x)dx$ is the electric field and energy dependent WKB factor, given by Eq. 3.18. The value of τ_o depends on the model used for the oxide trap potential and was shown to vary negligibly with electric field at the bias conditions of interest. The time for tunneling from the valence band $\tau_{ot/vb}$ is analogous.

In the case where the oxide trap is lower in energy than the conduction band state, with the excess energy dissipated by the multiple emission of phonons through a lattice relaxation process,

$$\tau_{ot/lr} = \left[\sigma_n \overline{v} e^{-\Delta E_{Bc}/kT}(n_s + n_1) + \sigma_p \overline{v} e^{-\Delta E_{Bv}/kT}\, e^{\lambda_v(x)/\lambda_c(x)}(p_s + p_1)\right]^{-1} e^{\lambda_c(x)} \qquad (3.99)$$

where the WKB integrals $\lambda_c(x)$ and $\lambda_v(x)$ are evaluated at the bottom of the silicon conduction band and the top of the valence band, respectively [Eqs. 3.28, 3.32], where the tunneling is assumed to occur. The values of σ_n and σ_p depend on the model used for the oxide trap potential.

Trap-to-Trap Transitions ($i = it$):

Interface state assisted tunneling is a second-order process characterized by two time constants:

$$\tau_{it} = [\sigma_{no}\overline{v}(n_s + n_1) + \sigma_{po}\overline{v}(p_s + p_1)]^{-1} \quad (3.100)$$

$$\tau_{ot/it} = \tau_o e^{\lambda(x)} \quad (3.101)$$

Values of τ_o for this mechanism were evaluated for several model oxide trap potentials. The WKB integral can be calculated from one of the two-band approximations presented in Sec. 3.6.1. However, it was shown that, due to the large effective mass and potential barrier confronted by holes, the problem can be treated to very good approximation by considering only electrons tunneling into the oxide conduction band. While a rigorous description demands that both bands be taken into account, we have not found this to be necessary in simulations on NMOS transistors, as described in Sec. 7.2.4, where the attenuation constant in the oxide is calculated using only the conduction band.

The dynamics of the two-step process were solved numerically for the case of a continuous spatial distribution of oxide traps [Eqs. 3.67–3.69]. This exact solution was compared with several approximate solutions. Since it is a second-order process, the general solution has the form

$$R_{it}(E, x, t) \equiv Ae^{-t/\tau_1} + Be^{-t/\tau_2}$$

where τ_1 and τ_2 are functions of $\tau_{ot/it}$ and τ_{it} [Eqs. 3.80, 3.82]. Under the assumption of widely separated time constants, which is usually the case, a first-order approximation can be made:

$$R_{it}(E, x, t) \equiv e^{-t/\tau_{ot}}$$

where [Eqs. 3.78, 3.83]

$$\tau_{ot} = \tau_{ot/it} + \tau_{it} \quad (3.102)$$

In strong inversion, when $\tau_{ot/it} \gg \tau_{it}$, the tunneling step becomes rate-limiting and $\tau_{ot} \approx \tau_{ot/it}$. However, for trap levels near midgap, in weak inversion or depletion, the thermal step may become rate limiting with the result that $\tau_{ot} \approx \tau_{it}$. This has an important effect on the temperature dependence of the threshold voltage shifts as discussed in the next chapter.

3.8 Discussion

The complete process of the charging and discharging of an oxide trap may involve any combination of the initial states, final states and tunneling mechanisms described above and indicated in Fig. 3.1(b). The rate at which a given oxide trap charges or

discharges is the superposition of all possible combinations of these channels operating serially and in parallel. The total number of possible paths is clearly too large to attempt to model in an exact manner. However, for channels which operate in parallel, a good approximation can be made by neglecting all but the most probable mechanisms. Since traps above the silicon conduction band edge are accessible to electron states at the same energy, elastic tunneling between these two sets of states is considered to be the dominant mechanism.

For trap energies below the conduction band, there are two competing mechanisms – the two-step interface state assisted process and the elastic tunneling/LRME process. These processes are superficially similar, and appear to be essentially identical in an energy band diagram [processes (5) and (6) in Fig. 3.1(b)]. Each involves an elastic tunneling step and a thermal (phonon-assisted) transition; the latter occuring at the interface in the two-step process and at the trap site in the case of LRME. Closer examination, however, reveals several important differences:

1. The two-step process involves an interaction between three physical entities — an inversion layer electron, a defect at the interface and a defect in the insulator. The LRME process involves just two – the inversion layer electron and the oxide trap. This is why the latter process can be described by a first-order differential equation while the former requires a second-order differential equation. This is clear in the energy band diagram of Fig. 3.1(b), where the ground and excited state of process (5) occupy the same location in space. There is just one defect, which can contain one electron, the difference in energy being caused by a different positioning of the surrounding lattice ions before and after capture.

2. The tunneling steps for the two processes occur at different energies and hence through different potential barriers. Since LRME can occur from energies above the conduction band while the two-step process occurs from states within the silicon bandgap, the wavefunctions of the latter are more strongly damped, decreasing the tunneling probability.

3. In LRME, the tunneling occurs from a delocalized conduction band state, while the two-step process occurs via a localized interface state. Since the inversion layer bathes the interface with electrons, while interface states occur randomly in the plane, electrons tunneling by the LRME mechanism have a higher probability of coming within tunneling distance of an oxide trap.

4. The thermal transitions for LRME and the two-step process occur in the insulator and in the silicon, respectively, where the phonon spectra differ.

Mathematically, these two processes are described by different time constants. The effective time constant for the two-step process [Eq. 3.102] looks like the sum of a tunneling term and a thermal term, while the LRME time constant [Eq. 3.99]

is proportional to the product of these terms. In order to determine how the two processes might by distinguished in an experiment, consider the time constants for an electron trap ($n_1 \gg p_1$) in two limiting cases:

1. $n_s \gg n_1$, $\tau_{ot/it} \gg \tau_{it}$ (e.g. strong inversion)

$$\text{Two-Step:} \quad \tau_{ot/it} = \tau_o e^{x/x_o}$$

$$\text{LRME:} \quad \tau_{ot/lr} = \frac{1}{\sigma_n \overline{v} n_s} e^{\Delta E_B/kT} e^{x/x_o} \sim \frac{1}{C_{ox}(V_{GS} - V_T)} e^{\Delta E_B/kT} e^{x/x_o}$$

2. $n_s \ll n_1$, $\tau_{ot/it} \ll \tau_{it}$ (e.g. weak inversion/depletion, traps near midgap)

$$\text{Two-Step:} \quad \tau_{ot/it} = \frac{1}{\sigma_{no} \overline{v} N_c} e^{(E_c - E)/kT}$$

$$\text{LRME:} \quad \tau_{ot/cb} = \frac{1}{\sigma_n \overline{v} N_c} e^{(E_c - E + \Delta E_B)/kT} e^{x/x_o}$$

While the two mechanisms are fundamentally different from a theoretical point of view, it will be shown in Chapters 4 and 7 that from an experimental point of view, they are very difficult to distinguish. Both time constants are dominated by the exponential spatial dependence in strong inversion and become thermally activated for deep traps in weak inversion and depletion. The major difference between the two occurs in inversion, when the prefactor of the exponential in the LRME time constant exhibits a direct proportionality to gate voltage and a thermal activation factor. These differences will be investigated further in Chapters 4 and 7.

Chapter 4

Threshold Voltage Relaxation in MOSFETs

4.1 Introduction

In this chapter, the theory developed in Chapter 3 is applied to the threshold voltage shifts which occur in MOSFETs when channel charge carriers tunnel to and from oxide traps under large-signal stress conditions.

In Chapter 1 it was shown that the application of a voltage pulse to the gate of a MOSFET causes a shift in the drain current, which decays slowly back to its original value when the pulse is removed. When referred to the input, this change can be expressed in terms of the *equivalent gate-source voltage shift* required to produce the same change in drain current. Section 4.2 considers the threshold voltage and mobility contributions to the equivalent gate-source voltage shift and shows that, because mobility fluctuations are *correlated* with the trapped charge, both can be described in terms of an *effective threshold voltage shift*. This reduces the problem to a threshold voltage calculation, to which the theory developed in Chapter 3 is directly applicable.

A general solution for the trapping-induced threshold voltage shift is derived in Sec. 4.3, which is exact within the approximations of the model developed in the previous chapter. When an electric field is applied to the gate of a MOSFET, the bending of the energy bands in the oxide causes trap energies to shift with respect to the interface. This effect, which was neglected in Chapter 3, requires two modifications of the theory. First, the value of the Fermi function at the interface, as seen by a given oxide trap, will depend on bias, as discussed in Sec. 4.3.1. Secondly, as a consequence of the change in trap energy levels during stress, the charging and discharging of a given trap may occur by different mechanisms. Section 4.3.2 introduces a formalism for the description of the various combinations of trap charging and discharging mechanisms and presents theoretical results for the contributions of

each of these components to the total threshold voltage shift.

Approximate solutions for the voltage and time dependence of the threshold voltage shift are developed in Sec. 4.4 for the case of strong inversion operation, which is the bias condition most commonly encountered in analog circuits. Temperature dependence is discussed in Sec. 4.5 and channel length and width dependence in Sec. 4.6. Section 4.7 discusses the manifestation of threshold voltage relaxation in the form of hysteresis effects in circuits, and considers several explanations for this behavior. Sec. 4.8 discusses some of the dynamic aspects of the threshold voltage relaxation effect which may occur under periodic stress waveforms in circuits. Finally, Sec. 4.9 summarizes the results of this chapter.

4.2 Formulation of the Problem

In the linear region, the drain current is given by

$$I_{DS} = \frac{W}{L}\mu_N Q_N V_{DS} \tag{4.1}$$

where Q_N is the total inversion layer charge (coul/cm^2),

$$Q_N = C_{ox}(V_{GS} - V_T) \tag{4.2}$$

Thus, a change in current may be produced by a change in mobility, a change in threshold voltage, or both. It is not possible, by our measurements, to experimentally distinguish these contributions.

The calculation of the equivalent gate-source voltage shift produced by a gate voltage pulse is a large-signal problem from the point of view of the band bending and carrier concentrations which occur between stress and relaxation. However, since the amount of charge which is trapped during stress is very small ($\sim 10^{10}$ coul/cm^2), changes in threshold voltage and mobility which occur during stress can be treated as small-signal perturbations from their pre-stress (nominal) values.

Assuming that V_{GS} is variable for reasons which will become clear below, the variation of the drain current $\Delta I_{DS}(E,x,t)$ with time, due to charge trapped at energy E and depth x is, by partial differentiation of Eq. 4.1,

$$\Delta I_{DS}(E,x,t) = \frac{I_{DS}}{\mu_N}\Delta\mu(E,x,t) + g_m \Delta V_{GS}(E,x,t) - g_m \Delta V_T(E,x,t) \tag{4.3}$$

where $g_m = \mu_N(W/L)C_{ox}V_{DS}$ is the transconductance. Note that each of incremental quantities in Eq. 4.3 depends on E and x through the trapped charge at that energy and depth.

The measurement circuit described in Chapter 6 directly measures the equivalent gate-source voltage shift required to keep I_{DS} constant. Solving for the change in

gate voltage required to make $\Delta I_{DS}(E,x,t) = 0$,

$$\Delta V_{GS}(E,x,t) = \Delta V_T(E,x,t) - (V_{GS} - V_T)\frac{\Delta\mu(E,x,t)}{\mu_N} \tag{4.4}$$

where V_{GS}, V_T and μ_N represent the nominal gate voltage, threshold voltage and mobility, respectively. Thus, the quantity that we measure consists of a threshold voltage shift plus a term proportional to the change in mobility.

4.2.1 Threshold Voltage Shift

Consider a charge $\Delta Q_{ot}(E,x,t)$ (coul/cm^2) in traps at energy E and localized at a depth x from the interface,

$$\Delta Q_{ot}(E,x,t) = -q\Delta n_{ot}(E,x,t)\Delta E\Delta x$$

The contribution to the threshold voltage shift $\Delta V_T(E,x,t)$ due to this trapped charge is the change in gate voltage required to produce the same surface potential and hence the same surface carrier concentrations, with $\Delta\mu = 0$. This requires that all of the trapped charge be imaged by an equal and opposite charge on the gate, which sets up an electric field in the oxide given by

$$\mathcal{E}(x') = \begin{cases} \frac{-\Delta Q_{ot}}{\epsilon_{ox}} & x < x' < t_{ox} \\ 0 & 0 < x' < x \end{cases} \tag{4.5}$$

Since the electric field due to this impulse of trapped charge is constant for $x < x' < t_{ox}$, the threshold voltage shift is just

$$\Delta V_T(E,x,t) = \frac{-\Delta Q_{ot}}{\epsilon_{ox}}(t_{ox} - x) \tag{4.6}$$

$$= \frac{q}{C_{ox}}\left(1 - \frac{x}{t_{ox}}\right)\Delta n_{ot}(E,x,t)\Delta E\Delta x \tag{4.7}$$

The total threshold voltage shift is obtained by integrating over all energy and space,

$$\Delta V_T(t) = \frac{q}{C_{ox}}\int_0^{t_{ox}}\int_E \left(1 - \frac{x}{t_{ox}}\right)\Delta n_{ot}(E,x,t)dEdx. \tag{4.8}$$

An alternative derivation of this result is given in Appendix E.

4.2.2 Correlated Mobility Fluctuations

Mobility fluctuations affect the equivalent gate-source voltage shift through the second term in Eq. 4.4. A trapped charge in the oxide causes a change in the mobility of inversion layer charge carriers due to increased coulombic scattering. This effect has

been demonstrated in 1/f noise measurements, where the change in channel conductance due to the removal of a channel charge carrier may be more or less than that expected, based on number fluctuations alone [152, 153].

The instantaneous mobility μ_n can be written, by Matthiessen's rule [127],

$$\frac{1}{\mu_n(E,x,t)} = \frac{1}{\mu_N} + \frac{1}{\mu_S(E,x,t)} \tag{4.9}$$

where $\mu_n(E,x,t)$ is the instantaneous mobility, μ_N is the average mobility, and $\mu_S(E,x,t)$ is the mobility due to coulombic scattering from trapped oxide charges at energy E and depth x.

Since number and mobility fluctuations arise from the same mechanism, mobility fluctuations can be correlated with the trapped charge [154]:

$$\frac{1}{\mu_S(E,x,t)} = S(x)q\Delta n_{ot}(E,x,t)\Delta E\Delta x \tag{4.10}$$

where $S(x)$ is the scattering rate (coul/V-s). From Eqs. 4.9 and 4.10, the mobility fluctuation is

$$\begin{aligned}\Delta\mu(E,x,t) &= \mu_n - \mu_N = -\frac{\mu_N^2 S(x)q\Delta n_{ot}(E,x,t)\Delta E\Delta x}{1+\mu_N S(x)q\Delta n_{ot}(E,x,t)\Delta E\Delta x} \\ &\approx -\mu_N^2 S(x)q\Delta n_{ot}(E,x,t)\Delta E\Delta x \end{aligned} \tag{4.11}$$

where the approximate form has assumed small fluctuations.

In general, the scattering rate depends on the distance of the trapped charge from the interface [155, 156]. Brews [155] has proposed an approximate expression of the form,

$$S(x) \approx S_o \ln\left[1+\left(\frac{r_{max}}{x}\right)^2\right]$$

where

$$r_{max} = \frac{\epsilon_{ox}+\epsilon_{Si}}{C_{ox}+C_s}$$

and C_s is the semiconductor capacitance. r_{max} has the interpretation as the distance, beyond which image terms cause the coulombic scattering potential to fall off more rapidly than $1/r$.

4.2.3 Equivalent Gate-Source Voltage Shift

Substituting Eqs. 4.7 and 4.11 into Eq. 4.4, the contribution to $\Delta V_{GS}(E,x,t)$ from traps at energy E and depth x is

$$\Delta V_{GS}(E,x,t) = \frac{q}{C_{ox}}\left[1-\frac{x}{t_{ox}}+\mu_N C_{ox}(V_{GS}-V_T)S(x)\right]\Delta n_{ot}(E,x,t)\Delta E\Delta x \tag{4.12}$$

and the total equivalent gate-source voltage shift is obtained by integrating over all energy and space:

$$\begin{aligned}\Delta V_{GS}(t) &= \int\int \Delta V_{GS}(E,x,t)dEdx \qquad (4.13)\\ &= \frac{q}{C_{ox}}\int_0^{t_{ox}}\int_E \left[1-\frac{x}{t_{ox}}+\mu_N C_{ox}(V_{GS}-V_T)S(x)\right]\Delta n_{ot}(E,x,t)\,dE\,dx\end{aligned}$$

When $S(x) \approx 0$, the equivalent gate-source voltage shift [Eq. 4.13] reduces to just a threshold voltage shift [Eq. 4.8].

An upper bound on $S(x)$ of about 3.5×10^3 V-s/coul has been estimated by Jayaraman [154] for charges located at the Si-SiO_2 interface. Consider an NMOS transistor with a 250Å gate oxide, so that $x \ll t_{ox}$. Using this value for S, with $V_{GS} - V_T = 0.5$V and $\mu_N = 700\text{cm}^2/\text{V-s}$, the correlated mobility term in Eq. 4.13 is 0.17, much less than the threshold voltage term. More recent work by Gross [118] indicates that 1/f noise data on MOSFETs can be fit by a scattering rate of the form

$$S(x) = \frac{S_o}{Q_N} = \frac{S_o}{C_{ox}(V_{GS}-V_T)} \qquad (4.14)$$

where $S_o \approx 0.02$ V-s/cm^2 for NMOS and is much smaller for PMOS. This result implies that the correlated mobility term may exceed the threshold voltage term in an NMOS transistor by an order of magnitude. It is evident from Eqs. 4.9 and 4.13 that correlated mobility fluctuations are more important the larger the mobility μ_N due to lattice scattering and, for this reason, are usually much smaller in PMOS transistors.

Using Eq. 4.14 in Eq. 4.13, the equivalent gate-source voltage shift can be written

$$\Delta V_{GS}(t) = \frac{q}{C_{ox}}\int_0^{t_{ox}}\int_E \left(1-\frac{x}{t_{ox}}+\mu_N S_o\right)\Delta n_{ot}(E,x,t)\,dE\,dx \qquad (4.15)$$

4.2.4 Effective Threshold Voltage Shift

With the change of variables,

$$\begin{aligned}t'_{ox} &= (1+\mu_N S_o)t_{ox}\\ C'_{ox} &= \epsilon_{ox}/t'_{ox}\end{aligned}$$

Equation 4.15 can be written

$$\Delta V'_T(t) = \frac{q}{C'_{ox}}\int_0^{t'_{ox}}\int_E \left(1-\frac{x}{t'_{ox}}\right)\Delta n_{ot}(E,x,t)\,dE\,dx \qquad (4.16)$$

where the upper limit of integration, which is effectively infinite for the small tunneling depths involved, has also been replaced by t'_{ox}. This result shows that the change in

drain current, which is produced by both mobility and threshold voltage fluctuations, can be viewed as being due to a threshold voltage shift only, with magnitude given by Eq 4.16, and a constant mobility. This *effective threshold voltage shift* is obtained by calculating the trapped charge in the usual manner, but in the transformation from charge to threshold voltage shift [Eq. 4.8], the oxide thickness is scaled by a factor $(1 + \mu_N S_o)$. For $x \ll t_{ox} < t'_{ox}$, which is satisfied for all of the devices used in this thesis, the effective threshold voltage shift is just the normal threshold voltage shift that would be calculated on the basis of the trapped charge alone, without mobility fluctuations, multiplied by a constant scale factor $(1 + \mu_N S_o)$.

The problem of calculating the equivalent gate-source voltage shift, at a given temperature, can therefore be treated formally as a threshold voltage calculation. For the remainder of this thesis, we will therefore disregard mobility fluctuations, with the understanding that they may affect the amplitude, but not the time or voltage dependence of any of the results.

There are two cases in which our neglect of the mobility fluctuation term may need to be reconsidered. First, since the mobility due to lattice scattering, μ_N, varies with temperature as $T^{-3/2}$, correlated mobility fluctuations can, if sufficiently large, cause a decrease in ΔV_{GS} with increasing temperature. Secondly, the trap distributions $D_{ot}(E, x)$ extracted from threshold voltage relaxation measurements [Sec. 7.2.4] will, in general, depend on the magnitude of the $\mu_N S_o$ term in Eq. 4.15.

4.3 General Solution

Having shown that mobility fluctuations can be incorporated into an effective threshold voltage shift, this section addresses the calculation of the time-dependent threshold voltage shift due to the various trap charging/discharging processes presented in Chapter 3.

For the purposes of the following discussion, it will be assumed that the traps are acceptor-like, i.e. neutral when empty and negatively charged when filled, which is known to be the case for interface states in the upper half of the silicon bandgap [157]. In fact, the threshold voltage shift is independent of the charge character of the traps because Eq. 4.6 depends only on the change in trapped charge ΔQ_{ot}. For example, under positive gate voltage stress, initially empty acceptor (donor) traps become filled, changing from neutral (positive) to negative (neutral). In either case, $\Delta Q_{ot} < 0$. Similarly, for a negative gate voltage, $\Delta Q_{ot} > 0$ regardless of the type of trap.

4.3.1 The Effect of Band Bending in the Oxide

In the previous chapter, all results were written in terms of the energy E of the tunneling electron. In this chapter we are interested in following the occupancy of a

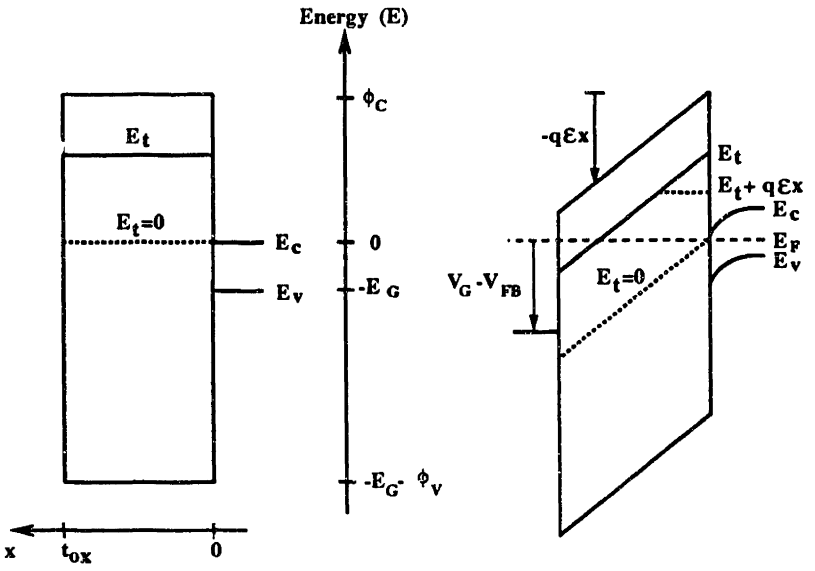


Figure 4.1: Energy band diagram of the MOS structure showing definition of coordinates for measurement of oxide trap energies; (a) flatband (b) with applied electric field $\mathcal{E}$.

given oxide trap with time, which requires keeping track of the mechanisms by which each trap charges and discharges. It is convenient for this purpose, to express all energies in terms of the energy E_t of a trap in a reference system fixed with respect to the oxide band structure.

The choice of a particular reference level as the zero of energy is arbitrary, however it is convenient for our purposes to choose a common zero reference for the entire system. Since we have assumed electronic energies to be measured upward from the silicon conduction band, we shall take the bottom of the silicon conduction band at the interface ($x = 0$) as the zero of energy for the system. This reference level extends horizontally through the oxide under flatband conditions as shown in Fig. 4.1(a). Another common convention is to specify electron traps by their depth below the insulator conduction band and hole traps by their energy above the insulator valence band. It is a simple matter to convert from one convention to the other, however, the convention used here has the advantage providing a consistent energy reference for the entire system.

The ground state energy of a trap is a chemical property which remains invariant with respect to the oxide band structure. In the band bending approximation, when an electric field $\mathcal{E}$ is applied across the oxide, all trap energy levels simply shift rigidly by an amount $q\mathcal{E}x$ with respect to energies at the interface, where $q = 1.602 \times 10^{-19}$ coulomb is the electron charge. Note that, due to our choice of the x direction to the left, a positive voltage applied to the gate produces a negative electric field $\mathcal{E} < 0$ as

shown in Fig. 4.1(b), so that trap energies are lowered with respect to the interface by an amount $-q\mathcal{E}x > 0$. In order to follow the occupancy with time of a particular oxide trap, the reference level for measuring trap energies must remain fixed with respect to the oxide conduction and valence bands and must therefore, also shift under an applied field as shown in Fig. 4.1(b). Thus, an oxide trap density $D_{ot}(E_t, x)$ with ground state energy $E_t = 0$, always lies $\phi_c = 3.1$ eV below the oxide conduction band regardless of bias conditions.

Since we are assuming elastic tunneling, the energy E of an electron at $x = 0$ is related to the energy level E_t of a trap at x into which it tunnels by

$$E = E_t + q\mathcal{E}x \tag{4.17}$$

as shown in Fig. 4.1(b). For example, an electron at the bottom of the conduction band in the semiconductor ($E = 0$) can tunnel into a trap at x of energy $E_t = -q\mathcal{E}x$, i.e. at energy $\phi_c + q\mathcal{E}x$ below the oxide conduction band. Note that the tunneling, which is elastic, occurs between states at the same energy. The inequivalence of E_t and E is a consequence of our choice of a fixed reference level for the measurement of electron energies at the interface, while the reference level for measuring trap energies $D_{ot}(E_t, x)$ shifts under an applied electric field.

The results of Chapter 3, which are written in terms of the energy of the tunneling electron, can be written in terms of the trap energy through the transformation (4.17). For brevity, we shall occassionally suppress the explicit dependence of the Fermi factors and time constants on energy and depth and use the sh thand notation

$$f \equiv f(E_t + q\mathcal{E}x) = \left\{1 + \exp\left[\frac{E_t + q\mathcal{E}x - E_F}{kT}\right]\right\}^{-1} \tag{4.18}$$

$$f_0 \equiv f_0(E_t + q\mathcal{E}_o x) = \left\{1 + \exp\left[\frac{E_t + q\mathcal{E}_o x - E_{F0}}{kT}\right]\right\}^{-1} \tag{4.19}$$

$$\tau_{ot/i} \equiv \tau_{ot/i}(E_t + q\mathcal{E}x, x) \tag{4.20}$$

$$\tau^o_{ot/i} \equiv \tau^o_{ot/i}(E_t + q\mathcal{E}_o x, x) \tag{4.21}$$

$$\tau_{it} \equiv \tau_{it}(E_t + q\mathcal{E}x) \tag{4.22}$$

$$\tau^o_{it} \equiv \tau^o_{it}(E_t + q\mathcal{E}_o x) \tag{4.23}$$

where $\mathcal{E}_o$ and E_{F0} are the electric field in the oxide and the Fermi level for electrons at the interface before stress, and $\mathcal{E}$ and E_F are the corresponding quantities during stress. Explicit expressions for the time constants as a function of E and x have been derived in the previous chapter.

4.3.2 Trapping Components

In Chapter 3, four basic mechanisms were considered by which oxide traps could be charged and discharged: (1) direct elastic tunneling transitions of electrons between oxide traps and silicon conduction band (cb) states, (2) direct elastic tunneling

transitions of holes between oxide traps and silicon valence band (vb) states, (2) the trap-assisted tunneling process in which electrons tunnel between oxide traps and fast surface states and from there communicate with the silicon bands by a thermal process (it), and (3) elastic tunneling into an excited state of an oxide trap, followed by lattice relaxation multiphonon emission (lr). Since the bending of the oxide bands during stress shifts trap energies with respect to the interface, it is possible for traps which charge by one of these mechanisms to discharge by another.

During stress, a trap at energy E_t and depth x may be filled by an electron which tunnels via process i. Following stress, this trapped electron may tunnel out by process j, where i and j designate the processes described above (cb, vb, it, lr). Each oxide trap, which is specified by its ground state energy E_t and depth x from the interface, can be uniquely labeled by i/j, where i designates the process by which the oxide trap charges during stress, and j represents the process by which the trap discharges following stress. For the four mechanisms that we have considered, there are 16 possible combinations of charging and discharging processes for a given oxide trap. The 9 possible combinations due to the direct and two-step processes (cb, vb, it) are shown schematically in Fig. 4.2. These diagrams catalog the possible transitions and are not meant to depict energy or spatial relationships. Thermal transitions are shown by dotted lines and tunneling transitions by solid lines. The notation used for the time constants during stress ($\tau_{ot/i}$) and during relaxation ($\tau^o_{ot/j}$) are also shown. Three of the mechanisms in Fig. 4.2 occur only for positive gate voltage pulses (it/cb, vb/it, vb/cb), three occur only for negative gate voltages pulses (cb/it, cb/vb, it/vb) while the other three (cb/cb, it/it, vb/vb) occur for any stress voltage.

The addition of lattice relaxation (lr) introduces 7 additional combinations (lr/cb, cb/lr, lr/lr, lr/it, it/lr, lr/vb, vb/lr). Three examples are illustrated in Fig. 4.3, where excited trap energy levels are represented by dashed circles. It is also possible for an electron or hole to tunnel elastically from an interface state into an oxide trap, followed by a lattice relaxation. This possibility is not considered in detail in this thesis, but may easily be incorporated into the general framework.

The total threshold voltage shift is the superposition of the contributions due to each of these components. This section is devoted to the derivation of general expressions for each of these trapping components.

In Chapter 3 it was shown that the non-steady-state occupancy function for oxide traps which capture electrons by tunneling from states of type $i \in \{cb, vb, it, lr\}$ was of the form

$$f_{ot}(E,x,t) = f_{ot}(E,x,\infty) + [f_{ot}(E,x,0^-) - f_{ot}(E,x,\infty)]R_i(E,x,t) \qquad (4.24)$$

where the function $f_{ot}(E,x,t)$ is the non-steady-state occupancy function for an oxide trap at energy E and depth x, and $f_{ot}(E,x,0)$ and $f_{ot}(E,x,\infty)$ denote its initial and final values. $R_i(E,x,t)$ is a generalized notation for the relaxation transient due to the charging of an oxide trap by mechanism i, which may represent either a first-order

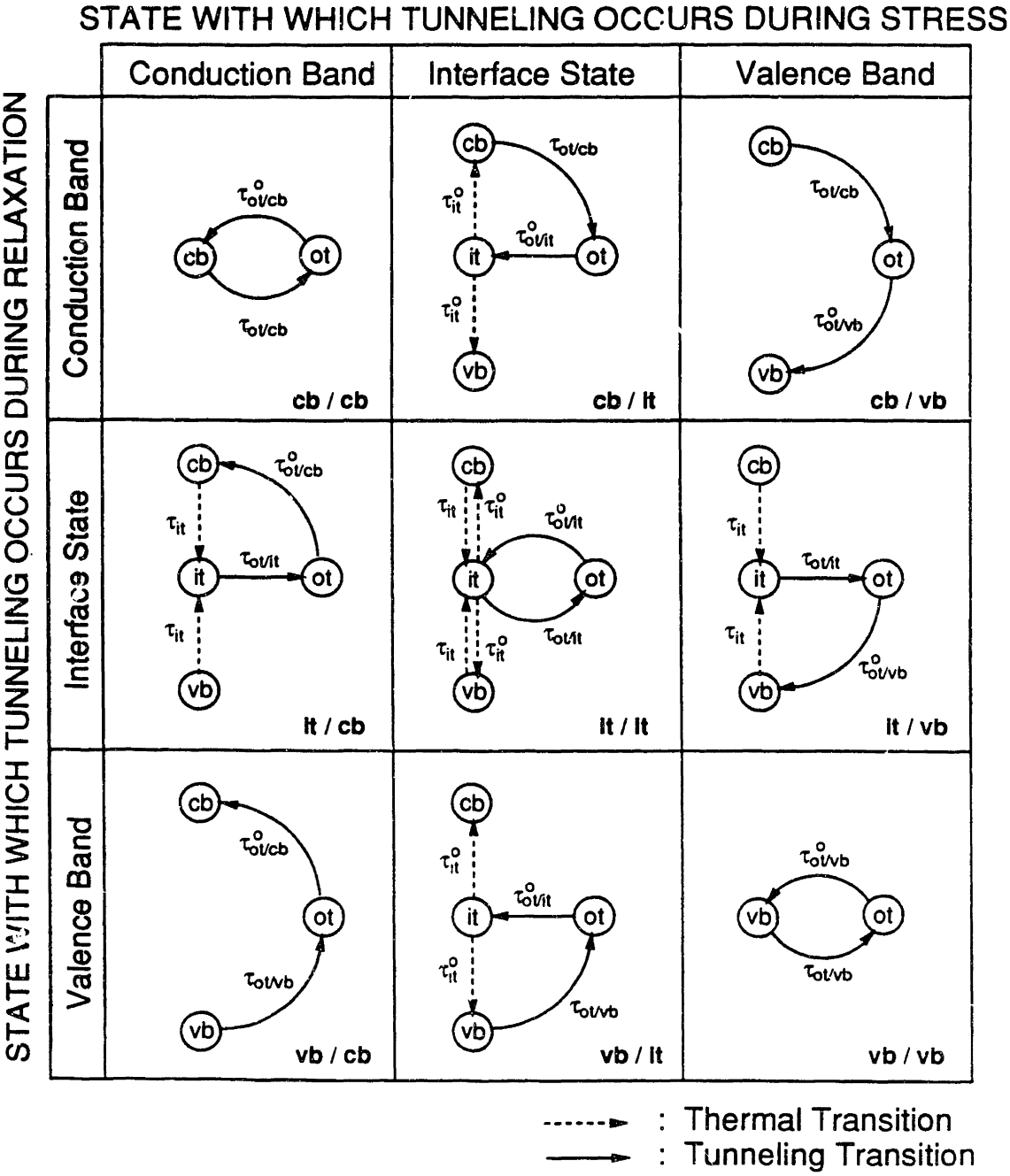


Figure 4.2: The nine possible combinations of direct and interface state assisted tunneling transitions by which an oxide trap can communicate with the silicon bands during a pulsed gate voltage measurement.

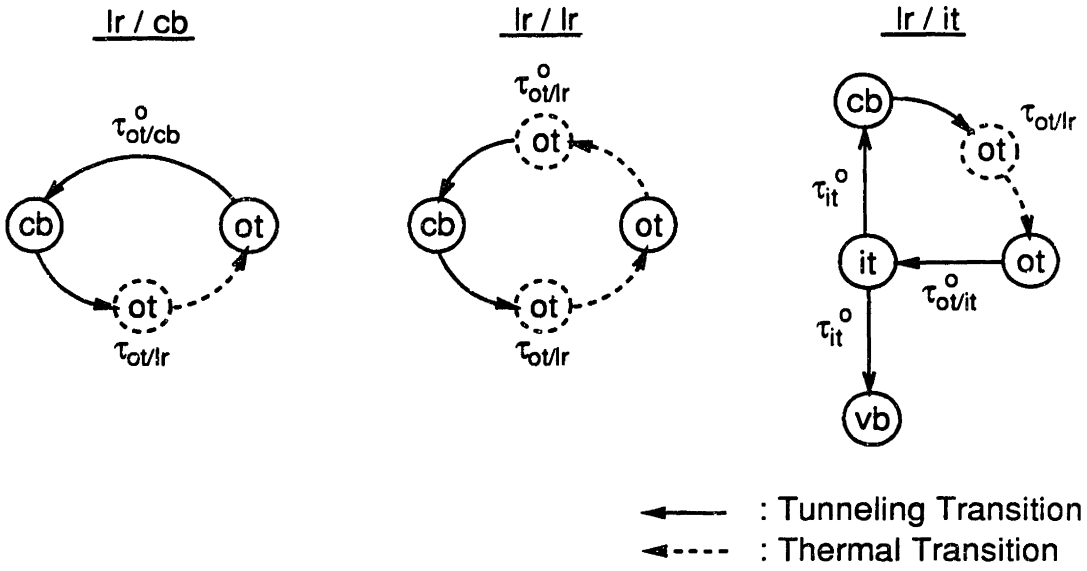


Figure 4.3: Examples of additional tunneling channels made possible by lattice relaxation multiphonon emission.

or a second-order response. From Chapter 3,

$$
\begin{aligned}
R_{cb}(E,x,t) &\equiv e^{-t/\tau_{ot/cb}(E,x)} \\
R_{vb}(E,x,t) &\equiv e^{-t/\tau_{ot/vb}(E,x)} \\
R_{it}(E,x,t) &\equiv \left[\frac{\tau_{ot/it}(E,x)}{\tau_{ot/it}(E,x)-\tau_{it}(E)}\right] e^{-t/\tau_{ot/it}(E,x)} - \left[\frac{\tau_{it}(E)}{\tau_{ot/it}(E,x)-\tau_{it}(E)}\right] e^{-t/\tau_{it}(E)} \\
R_{lr}(E,x,t) &= e^{-t/\tau_{ot/lr}(E,x)}
\end{aligned}
$$

where the partially coupled approximation (Approximation 2) of Sec. 3.5.2 has been used for R_{it}. Any of the other approximations derived in that section may also be used depending on the required accuracy.

Consider first the time interval during stress, $0 < t < t_s$, when traps are charging. Expressing all quantities in terms of the trap energy using Eq. 4.17, the initial and final values of the trap occupancy function are determined by the values of the Fermi function at the interface before and during stress, as seen by an oxide trap at energy E_t,

$$f_{ot}(E_t,x,0^-) = f_0(E_t+q\mathcal{E}_o x) \quad (4.25)$$

$$f_{ot}(E_t,x,\infty) = f(E_t+q\mathcal{E}x) \quad (4.26)$$

where we have used the assumption that equilibrium prevails before, and a very long time after, the applied stress pulse. Thus, for $0 < t < t_s$,

$$f_{ot}(E_t,x,t) = f(E_t+q\mathcal{E}x) + [f_0(E_t+q\mathcal{E}_o x) - f(E_t+q\mathcal{E}x)]R_i(E_t+q\mathcal{E}x,x,t) \quad (4.27)$$

where $R_i(E,x,t)$ is also transformed by Eq. 4.17 since it was derived in terms of the energy of the tunneling electron.

When the stress voltage is removed, the device returns to its pre-stress bias condition. In terms of the variable $t' = t - t_s$, referenced to the end of the stress interval, the initial and final values of the trap occupancy function are

$$f_{ot}(E_t, x, t' = 0^-) = f_{ot}(E_t, x, t = t_s) \quad (4.28)$$

$$f_{ot}(E_t, x, t' \to \infty) = f_0(E_t + q\mathcal{E}_o x) \quad (4.29)$$

Thus, the non-steady-state occupancy function for traps which fill by mechanism i and empty by mechanism j is, using Eqs. 4.24 and 4.28–4.29,

$$f_{ot}(E_t, x, t) = f_0(E_t + q\mathcal{E}_o x) + [f_{ot}(E_t, x, t_s) - f_0(E_t + q\mathcal{E}_o x)] R_j^o(E_t + q\mathcal{E}_o x, x, t - t_s) \quad (4.30)$$

where $R_j^o(E,x,t)$ is the relaxation transient at the post-stress bias condition, which is the same as the pre-stress bias condition:

$$R_{cb}^o(E,x,t) \equiv e^{-t/\tau_{ot/cb}^o(E,x)}$$

$$R_{vb}^o(E,x,t) \equiv e^{-t/\tau_{ot/vb}^o(E,x)}$$

$$R_{it}^o(E,x,t) \equiv \left(\frac{\tau_{ot/it}^o(E,x)}{\tau_{ot/it}^o(E,x) - \tau_{it}^o(E)}\right) e^{-t/\tau_{ot/it}^o(E,x)} - \left(\frac{\tau_{it}^o(E)}{\tau_{ot/it}^o(E,x) - \tau_{it}^o(E)}\right) e^{-t/\tau_{it}^o(E)}$$

$$R_{lr}^o(E,x,t) = e^{-t/\tau_{ot/lr}^o(E,x)}$$

The change in trap occupancy that occurs between $t = 0$ and t ($t > t_s$) for traps in category i/j is therefore,

$$\Delta f_{ot_{i/j}}(E_t, x, t) = f_{ot}(E_t, x, t) - f_{ot}(E_t, x, 0^-) = f_{ot}(E_t, x, t) - f_0(E_t + q\mathcal{E}_o x) \quad (4.31)$$

Combining Eqs. 4.27, 4.30 and 4.31, the change in oxide trap occupancy is

$$\Delta f_{ot_{i/j}}(E_t, x, t) = [f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x)] \times [1 - R_i(E_t + q\mathcal{E}x, x, t_s)] R_j^o(E_t + q\mathcal{E}_o x, x, t - t_s) \quad (4.32)$$

In a similar manner, the change in the interface state occupancy function is

$$\Delta f_{it}(E_t, t) = [f(E_t) - f_0(E_t)](1 - e^{-t_s/\tau_{it}}) e^{-t_r/\tau_{it}} \quad (4.33)$$

Two types of measurements are performed, as described in Chap. 6. In the first type of measurement, the "initial" amplitude of the threshold voltage shift after a stress time t_s is monitored as a function of bias conditions and device characteristics. However, there is always a delay due to the finite response time of the test circuit so

that what is actually measured is not really the initial value at $t = t_s$, but the value of the threshold voltage shift after a stress time t_s followed by a short delay time $t_r = t - t_s$. This is important because, as will be shown in Sec. 4.7, the decay rate for oxide traps very close to the interface occurs in times shorter than the settling time of the measurement circuit and these traps therefore do not contribute to the measured response. Thus, the initial threshold voltage shift, $\Delta V_T(t = t_s)$, may be very much larger than indicated by the first time point which can actually be measured. In this type of experiment the relaxation time t_r is fixed while the stress time t_s is variable. In the second type of measurement, it is the relaxation of the threshold voltage shift as a function of time following the stress period which is of interest. Thus, the stress time t_s is fixed and the relaxation time t_r is variable. The point is that, in either case, the trap occupancy is always observed after some of the traps have been allowed to discharge, a fact which has been overlooked in previous studies. It is always the case that $t > t_s$ in any measurement and thus, the trap occupancy function depends on two time variables; the stress time t_s and the relaxation time $t_r = t - t_s$.

The total threshold voltage shift is due to the charge trapped in interface states and oxide traps:

$$\Delta V_T(t_s, t_r) = \Delta V_T(t_s, t_r)|_{it} + \Delta V_T(t_s, t_r)|_{ot} \tag{4.34}$$

where the interface state contribution is obtained by integrating Eq. 4.33 over all energies in the silicon bandgap

$$\Delta V_T(t_s, t_r)|_{it} = \frac{q t_{ox}}{\epsilon_{ox}} \int_{-E_G}^{0} [f(E) - f_0(E)] D_{it}(E) \left[1 - e^{-t_s/\tau_{it}(E)}\right] e^{-t_r/\tau_{it}^o(E)} dE \tag{4.35}$$

The oxide trap contribution is given by

$$\Delta V_T(t_s, t_r)|_{ot} = \sum_{i,j} \Delta V_{T(i/j)}(t_s, t_r) \tag{4.36}$$

where, using Eq. 4.32, the i/j component is given by

$$\begin{aligned} \Delta V_{Ti/j}(t_s, t_r) \;=\; & \frac{q}{\epsilon_{ox}} \int_0^{t_{ox}} \int_{E_1(i,j)}^{E_2(i,j)} (t_{ox} - x) [f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x)] \\ & \times [1 - R_i(E_t + q\mathcal{E}x, x, t_s)] R_j^o(E_t + q\mathcal{E}_o x, x, t_r) D_{ot}(E_t, x) dE_t dx \end{aligned} \tag{4.37}$$

The energy limits E_1 and E_2 in Eq. 4.37 depend on the tunneling mechanisms, i and j during stress and emission. The limits on x may also be subdivided into several intervals depending on bias. These considerations are addressed in the following section for the case of an NMOS transistor subjected to positive and negative gate voltage pulses. The PMOS case follows by analogy.

The formulation is completely general, with the differences between processes contained entirely in the relaxation transients R_i and the time constants $\tau_{ot/i}$. In the following section it will therefore be assumed, without loss of generality, that oxide

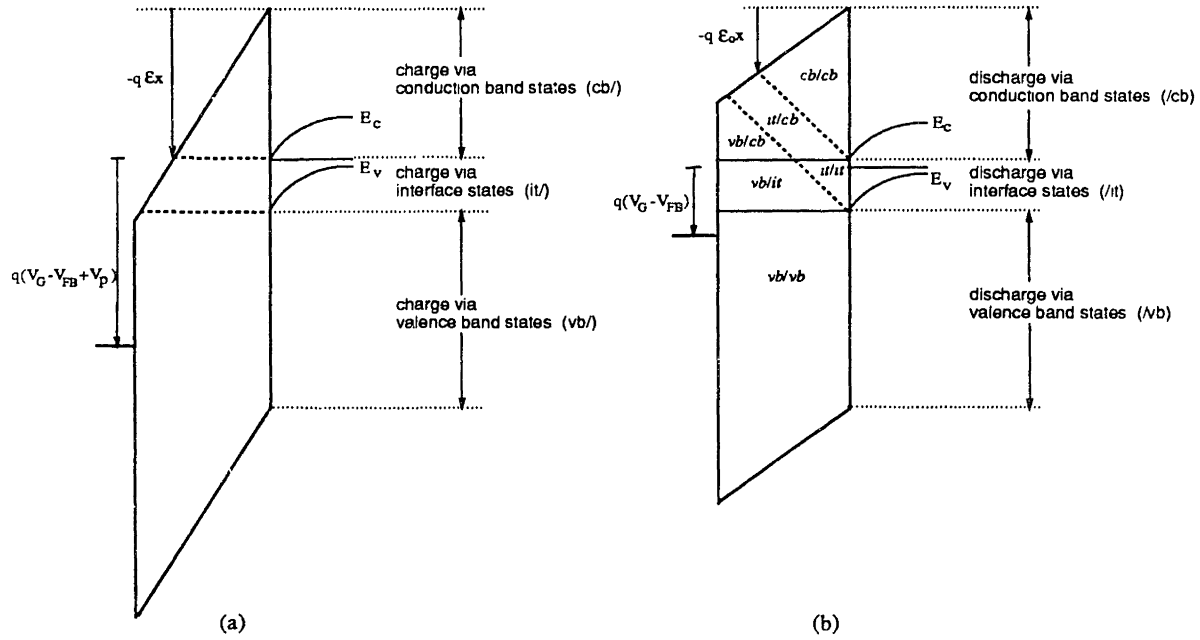


Figure 4.4: Band diagrams for an NMOS transistor subjected to positive gate voltage pulses, $V_p > 0$. (a) During stress, showing mechanism by which traps charge (b) After stress, showing mechanism by which traps empty, and classification of trap energy levels by fill mode/empty mode.

traps with energy levels falling within the silicon bandgap are filled by the two-step process (it). The lattice relaxation case follows by analogy with the substitution $it \longrightarrow lr$.

Positive Stress

Since we are interested in analog circuit applications, the MOSFET, prior to stress, is assumed to be biased in inversion, with $V_G' > 0$, where $V_G' = V_G - V_{FB}$. During stress with a gate voltage pulse $V_p = \Delta V_G > 0$, the oxide bands bend as shown in Fig. 4.4(a). A trap at distance x from the interface is lowered by an amount $-q\mathcal{E}x$, where $\mathcal{E} = -(V_G' + V_p)/t_{ox}$. From Fig. 4.17, traps at energies above the conduction band at the interface, $E = E_t + q\mathcal{E}x > E_c$ are then accessible to electrons tunneling elastically from conduction band states. Traps with $E_c - E_G < E_t + q\mathcal{E}x < E_c$ can be charged by tunneling from fast surface states and traps with $E_t + q\mathcal{E}x < E_c - E_G$ can fill by tunneling from valence band states. The charging process can therefore be categorized by the three types of initial states, which divide the oxide into three mutually exclusive regions bounded by the trap energies $E_t(x) = E_c - q\mathcal{E}x$ and $E_t(x) = E_c - E_G - q\mathcal{E}x$ as shown in Fig. 4.4(a).

Following stress, the gate voltage returns to its initial value. It is assumed that

Table 4.1: Trapping components for positive stress and $x < x_1$

Trapping Component	E_1	Stress Transient	E_2	Emission Transient
cb/cb	$E_c - q\mathcal{E}x$	$1 - R_{cb}$	$E_c + \phi_c$	R^o_{cb}
it/cb	$E_c - q\mathcal{E}_o x$	$1 - R_{it}$	$E_c - q\mathcal{E}x$	R^o_{cb}
it/it	$E_c - E_G - q\mathcal{E}x$	$1 - R_{it}$	$E_c - q\mathcal{E}_o x$	R^o_{it}
vb/it	$E_c - E_G - q\mathcal{E}_o x$	$1 - R_{vb}$	$E_c - E_G - q\mathcal{E}x$	R^o_{it}
vb/vb	$E_c - E_G - \phi_v$	$1 - R_{vb}$	$E_c - E_G - q\mathcal{E}_o x$	R^o_{vb}

the electric field in the oxide also returns to its pre-stress value. However, due to the presence of trapped charge near the interface, the band bending in the oxide after stress will differ somewhat from its pre-stress condition. This effect is assumed to be negligible due to the small magnitude of the trapped charge, as confirmed by the small amplitudes of the observed threshold voltage shifts. The inclusion of this effect would greatly complicate the formulation of the problem and cannot be justified by the small increase in accuracy.

When the stress is removed, a trap at x is shifted upward in energy by an amount $q(\mathcal{E}_o - \mathcal{E})x$. Thus, each of the regions in Fig. 4.4(a) demarcated by dashed lines maps into the corresponding region in Fig. 4.4(b). Electrons in oxide traps with $E_t + q\mathcal{E}_o x > E_c$ can then be reemitted to conduction band states, traps with $E_c - E_G < E_t + q\mathcal{E}_o x < E_c$ can be reemitted to fast surface states and traps with $E_t + q\mathcal{E}_o x < E_c - E_G$ can reemit to valence band states.

The emission of trapped charge can therefore be categorized by the three types of final states, demarcated by the trap energies $E_t(x) = E_c - q\mathcal{E}_o x$ and $E_t(x) = E_c - E_G - q\mathcal{E}_o x$. Thus, the oxide can be divided into six mutually exclusive regions as shown in Fig. 4.5, according to the mode by which the traps charge during stress and discharge following stress. The six possible combinations shown in Figs. 4.4(b) and 4.5 are the only ones which can occur under positive stress. However, it is not necessary that all of these occur. For example, at biases $\mathcal{E}_o - \mathcal{E} < E_G/qt_{ox}$, the vb/cb component will not occur. These regions indicate the appropriate integration limits on E_t and x to be used in Eq. 4.37 for the components of the threshold voltage shift. Table 4.1 shows the integration limits valid for $x < x_1$ where

$$x_1 = E_G/[q(\mathcal{E}_o - \mathcal{E})] \tag{4.38}$$

For $x > x_1$ the it/it component no longer contributes and the vb/cb component comes into play for energies $E_c - q\mathcal{E}_o x < E_t < E_c - E_G - q\mathcal{E}x$. For a change in voltage $V_p = \Delta V_G = 5\text{V}$ across a $250\mathring{A}$ oxide, $x_1 \approx 50\mathring{A}$ which is too deep to be of much importance at the time scales of interest. At higher voltages or in thinner oxides, the vb/cb component may contribute. For example, MNOS memory devices are charged by electrons tunneling from the valence band under a high electric field

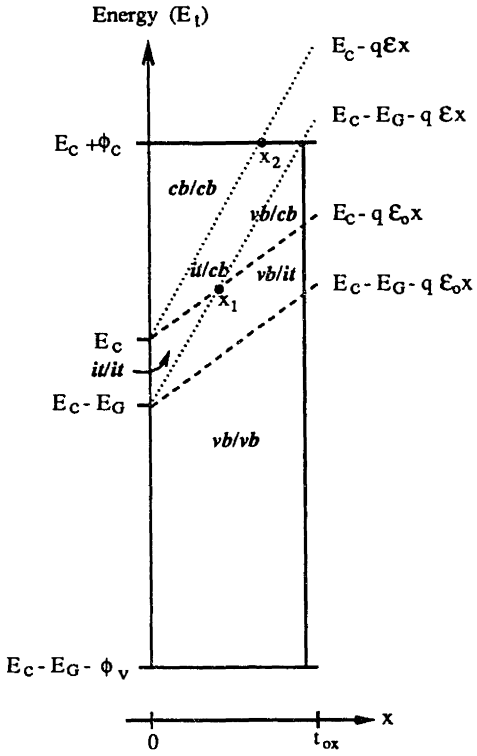


Figure 4.5: Classification of oxide trap energy levels into mutually exclusive regions in (E_t, x) space according to fill mode/empty mode for positive gate voltage pulses $V_p > 0$.

and discharged by the opposite polarity field which causes the trapped charge to be reemitted to empty conduction band states [90].

Table 4.2 shows the energy integration limits valid for the range $x_1 < x < x_2$ where

$$x_2 = \min\left(-\frac{\phi_c}{q\mathcal{E}}, t_{ox}\right) \tag{4.39}$$

For $\phi_c = 3$eV and a total voltage during stress of 7V across a 250$\mathring{A}$ oxide, $x_2 \approx 100\mathring{A}$ which is far beyond the tunneling depth during any stress of relevant duration. Various other special conditions on the energy limits arise, depending on bias conditions and trap depth. These bookkeeping details must be accounted for in the computer implementation of the equations, but do not contribute any new physical ideas and are important only under extreme bias conditions which are not of interest here. We are only interested in trapping in the near-interface region, on the order of 20$\mathring{A}$ from the interface for which Table 4.1 is valid under most bias conditions encountered in typical analog circuits. Due to the small depth actually accessed by tunneling during typical stress times, it is not necessary, in practice, to carry out the spatial integration

Table 4.2: Trapping components for positive stress and $x_1 < x < x_2$

Trapping Component	E_1	Stress Transient	E_2	Emission Transient
cb/cb	$E_c - q\mathcal{E}x$	$1 - R_{cb}$	$E_c + \phi_c$	R^o_{cb}
it/cb	$E_c - E_G - q\mathcal{E}x$	$1 - R_{it}$	$E_c - q\mathcal{E}x$	R^o_{cb}
vb/cb	$E_c - q\mathcal{E}_o x$	$1 - R_{vb}$	$E_c - E_G - q\mathcal{E}x$	R^o_{cb}
vb/it	$E_c - E_G - q\mathcal{E}_o x$	$1 - R_{vb}$	$E_c - e\mathcal{E}_o x$	R^o_{it}
vb/vb	$E_c - E_G - \phi_v$	$1 - R_{vb}$	$E_c - E_G - q\mathcal{E}_o x$	R^o_{vb}

over the entire thickness of the oxide.

In the case of a positive stress voltage, the total threshold voltage shift due to oxide traps is, from Eq. 4.36 and Fig. 4.5,

$$\begin{aligned}\Delta V_T &= \Delta V_{T\,cb/cb} + \Delta V_{T\,it/cb} + \Delta V_{T\,it/it} \\ &\quad + \Delta V_{T\,vb/it} + \Delta V_{T\,vb/cb} + \Delta V_{T\,vb/vb} \\ &\quad + \Delta V_{T\,lr/cb} + \Delta V_{T\,lr/lr} + \Delta V_{T\,vb/lr}\end{aligned} \tag{4.40}$$

where the contributions due to lattice relaxation have been included using the substitution $it \longrightarrow lr$.

Using Eq. 4.37 and the results in Tables 4.1 and 4.2, we can write the contribution of each component to the threshold voltage shift. Simulations indicate that the cb/cb, it/cb and it/it components dominate the threshold voltage shift while the other components are negligible under bias conditions of interest. We explicitly write out only these three dominant components for an NMOS device under positive stress. The other components follow by analogy. The results are only written for oxide depths $x < x_2$ which illustrates the procedure and is sufficient for the bias conditions of interest. The energy origin for oxide traps is taken at $E_c = 0$ as in Fig. 4.1. The E and x dependence of the Fermi factors and time constants have been suppressed for brevity and are given explicitly in Eqs. 4.18–4.23. Approximation 2 [Eq. 3.80] is used for the two-step process.

<u>cb/cb:</u>

$$\Delta V_{T\,cb/cb}(t_s, t_r) = \frac{q}{\epsilon_{ox}} \int_0^{x_2} \int_{-q\mathcal{E}x}^{\phi_c} (t_{ox} - x)(f - f_o) D_{ot}(E_t, x) \left[1 - e^{-t_s/\tau_{ot/cb}}\right] e^{-t_r/\tau^o_{ot/cb}} dE_t dx \tag{4.41}$$

<u>it/cb:</u>

$$\Delta V_{T\,it/cb}(t_s, t_r) = \frac{q}{\epsilon_{ox}} \int_0^{x_1} \int_{-q\mathcal{E}_o x}^{-q\mathcal{E}x} (t_{ox} - x)(f - f_o) D_{ot}(E_t, x) \tag{4.42}$$

Table 4.3: Trapping components for negative stress and $x < x_1$

Trapping Component	E_1	Stress Transient	E_2	Emission Transient
cb/cb	$E_c - q\mathcal{E}_o x$	$1 - R_{cb}$	$E_c + \phi_c$	R^o_{cb}
cb/it	$E_c - q\mathcal{E}x$	$1 - R_{cb}$	$E_c - q\mathcal{E}_o x$	R^o_{it}
it/it	$E_c - E_G - q\mathcal{E}_o x$	$1 - R_{it}$	$E_c - q\mathcal{E}x$	R^o_{it}
it/vb	$E_c - E_G - q\mathcal{E}x$	$1 - R_{it}$	$E_c - E_G - q\mathcal{E}_o x$	R^o_{vb}
vb/vb	$E_c - E_G - \phi_v$	$1 - R_{vb}$	$E_c - E_G - q\mathcal{E}x$	R^o_{vb}

$$\times \left[1 - \left(\frac{\tau_{ot/it}}{\tau_{ot/it} - \tau_{it}}\right) e^{-t_s/\tau_{ot/it}} + \left(\frac{\tau_{it}}{\tau_{ot/it} - \tau_{it}}\right) e^{-t_s/\tau_{it}}\right] e^{-t_r/\tau^o_{ot/cb}} dE_t dx$$

$$+\frac{q}{\epsilon_{ox}} \int_{x_1}^{x_2} \int_{-E_G - q\mathcal{E}x}^{-q\mathcal{E}x} (t_{ox} - x)(f - f_o) D_{ot}(E_t, x)$$

$$\times \left[1 - \left(\frac{\tau_{ot/it}}{\tau_{ot/it} - \tau_{it}}\right) e^{-t_s/\tau_{ot/it}} + \left(\frac{\tau_{it}}{\tau_{ot/it} - \tau_{it}}\right) e^{-t_s/\tau_{it}}\right] e^{-t_r/\tau^o_{ot/cb}} dE_t dx$$

it/it:

$$\Delta V_{Tit/it}(t_s, t_r) = \frac{q}{\epsilon_{ox}} \int_0^{x_1} \int_{-E_G - q\mathcal{E}_o x}^{-q\mathcal{E}_o x} (t_{ox} - x)(f - f_o) D_{ot}(E_t, x) \tag{4.43}$$

$$\times \left[1 - \left(\frac{\tau_{ot/it}}{\tau_{ot/it} - \tau_{it}}\right) e^{-t_s/\tau_{ot/it}} + \left(\frac{\tau_{it}}{\tau_{ot/it} - \tau_{it}}\right) e^{-t_s/\tau_{it}}\right]$$

$$\times \left[\left(\frac{\tau^o_{ot/it}}{\tau^o_{ot/it} - \tau^o_{it}}\right) e^{-t_r/\tau^o_{ot/it}} - \left(\frac{\tau^o_{it}}{\tau^o_{ot/it} - \tau^o_{it}}\right) e^{-t_r/\tau^o_{it}}\right] dE_t dx$$

Negative Stress

For negative gate voltage pulses $V_p < 0$, the electric field satisfies $\mathcal{E} > \mathcal{E}_o$. Figures 4.6(a) and (b) show the band diagrams during and after stress while Fig. 4.7 shows the classification of trap energy levels by fill mode/empty mode. The corresponding energy integration limits for Eq. 4.37 are summarized in Tables 4.3 and 4.4. In this case

$$x_1 = \frac{E_G}{q(\mathcal{E} - \mathcal{E}_o)} \tag{4.44}$$

$$x_2 = \min\left(\frac{\phi_v}{q\mathcal{E}}, t_{ox}\right) \tag{4.45}$$

Including lattice relaxation through the substitution $it \longrightarrow lr$, the total threshold

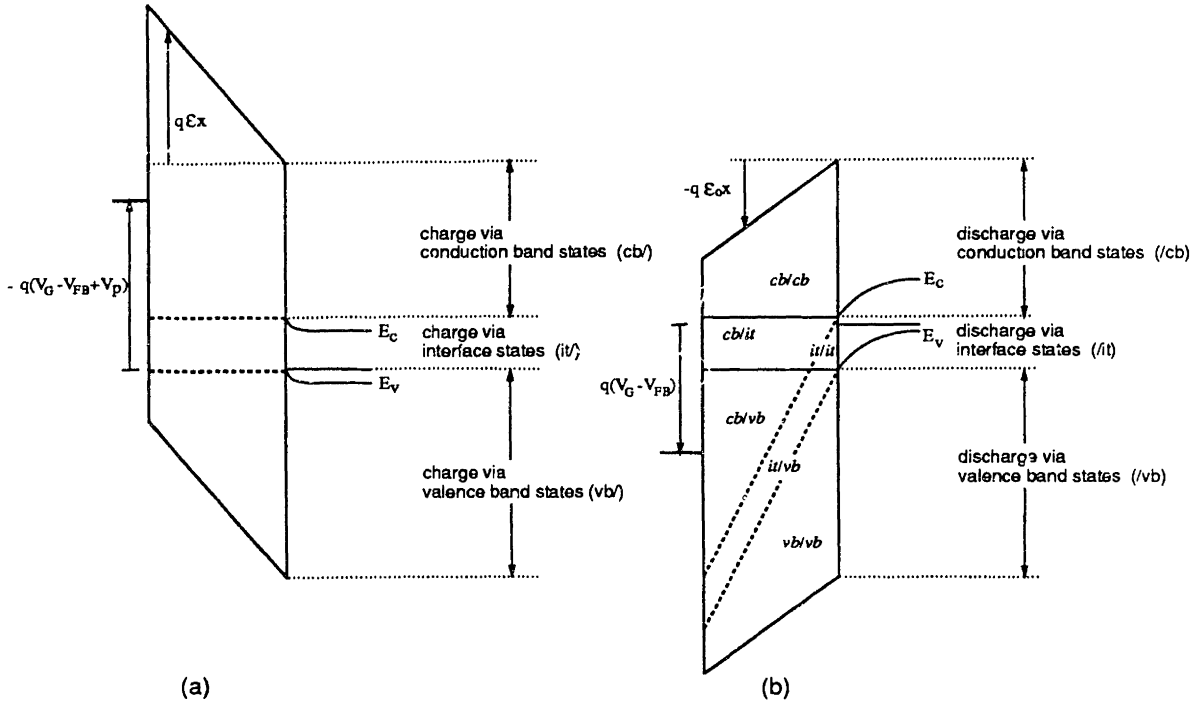


Figure 4.6: Band diagrams for an NMOS transistor subjected to negative gate voltage pulses, $V_p < 0$. (a) During stress, showing mechanisms by which traps charge (b) After stress, showing classification of trap energy levels by fill mode/empty mode.

voltage shift for negative voltages is again given by Eq. 4.36 as

$$\begin{aligned}\Delta V_T &= \Delta V_{Tcb/cb} + \Delta V_{Tcb/it} + \Delta V_{Tit/it} \\ &\quad + \Delta V_{Tcb/vb} + \Delta V_{Tit/vb} + \Delta V_{Tvb/vb} \\ &\quad + \Delta V_{Tcb/lr} + \Delta V_{Tlr/lr} + \Delta V_{Tlr/vb}\end{aligned} \tag{4.46}$$

where the trapping components are again given by Eq. 4.37 with the energy limits given in Tables 4.3 and 4.4 and the stress and relaxation transients as defined above. The three dominant components are written explicitly below.

cb/cb:

$$\Delta V_{Tcb/cb}(t_s, t_r) = \frac{q}{\epsilon_{ox}} \int_0^{x_2} \int_{-q\mathcal{E}_{ox}}^{\phi_c} (t_{ox} - x)(f - f_o) D_{ot}(E_t, x) \left[1 - e^{-t_s/\tau_{ot/cb}}\right] e^{-t_r/\tau^o_{ot/cb}} dE_t dx \tag{4.47}$$

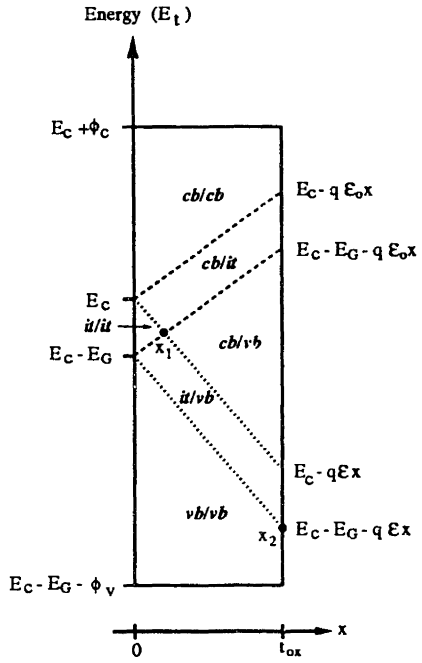


Figure 4.7: Classification of oxide trap energy levels into mutually exclusive regions in (E_t, x) space according to fill mode/empty mode for negative gate voltage pulses $V_p < 0$.

cb/it:

$$\Delta V_{Tcb/it}(t_s, t_r) = \frac{q}{\epsilon_{ox}} \int_0^{x_1} \int_{-q\mathcal{E}x}^{-q\mathcal{E}_o x} (t_{ox} - x)(f - f_o) D_{ot}(E_t, x)(1 - e^{-t_s/\tau_{ot/cb}}) \times \left[\left(\frac{\tau^o_{ot/it}}{\tau^o_{ot/it} - \tau^o_{it}} \right) e^{-t_s/\tau^o_{ot/it}} - \left(\frac{\tau^o_{it}}{\tau^o_{ot/it} - \tau^o_{it}} \right) e^{-t_s/\tau^o_{it}} \right] dE_t dx \quad (4.48)$$

Table 4.4: Trapping components for negative stress and $x_1 < x < x_2$

Trapping Component	E_1	Stress Transient	E_2	Emission Transient
cb/cb	$E_c - q\mathcal{E}_o x$	$1 - R_{cb}$	$E_c + \phi_c$	R^o_{cb}
cb/it	$E_c - E_G - q\mathcal{E}_o x$	$1 - R_{cb}$	$E_c - q\mathcal{E}_o x$	R^o_{it}
cb/vb	$E_c - q\mathcal{E}x$	$1 - R_{cb}$	$E_c - E_G - q\mathcal{E}_o x$	R^o_{vb}
it/vb	$E_c - E_G - q\mathcal{E}x$	$1 - R_{it}$	$E_c - q\mathcal{E}x$	R^o_{vb}
vb/vb	$E_c - E_G - \phi_v$	$1 - R_{vb}$	$E_c - E_G - q\mathcal{E}x$	R^o_{vb}

$$+\frac{q}{\epsilon_{ox}}\int_{x_1}^{x_2}\int_{-E_G-q\mathcal{E}_o x}^{-q\mathcal{E}_o x}(t_{ox}-x)(f-f_o)D_{ot}(E_t,x)(1-e^{-t_s/\tau_{ot/cb}})$$
$$\times\left[\left(\frac{\tau^o_{ot/it}}{\tau^o_{ot/it}-\tau^o_{it}}\right)e^{-t_r/\tau^o_{ot/it}}-\left(\frac{\tau^o_{it}}{\tau^o_{ot/it}-\tau^o_{it}}\right)e^{-t_r/\tau^o_{it}}\right]dE_t dx$$

it/it:

$$\Delta V_{Tit/it}(t_s,t_r) = \frac{q}{\epsilon_{ox}}\int_0^{x_1}\int_{-E_G-q\mathcal{E}_o x}^{-q\mathcal{E}x}(t_{ox}-x)(f-f_o)D_{ot}(E_t,x) \quad (4.49)$$
$$\times\left[1-\left(\frac{\tau_{ot/it}}{\tau_{ot/it}-\tau_{it}}\right)e^{-t_s/\tau_{ot/it}}+\left(\frac{\tau_{it}}{\tau_{ot/it}-\tau_{it}}\right)e^{-t_s/\tau_{it}}\right]$$
$$\times\left[\left(\frac{\tau^o_{ot/it}}{\tau^o_{ot/it}-\tau^o_{it}}\right)e^{-t_r/\tau^o_{ot/it}}-\left(\frac{\tau^o_{it}}{\tau^o_{ot/it}-\tau^o_{it}}\right)e^{-t_r/\tau^o_{it}}\right]dE_t dx$$

4.4 Approximate Voltage and Time Dependence

The complexity of Eqs. 4.40 and 4.46, with trapping components given by Eqs. 4.41–4.43 and 4.47–4.49 makes an exact closed form solution for the threshold voltage relaxation impractical. The integrations cannot be carried out in closed form for arbitrary oxide trap distributions in energy and depth. Furthermore, an exact solution requires the Fermi level, or equivalently, the surface potential, which cannot be expressed explicitly as a function of gate voltage. The situation is further complicated by the fact that the applied pulses drive the Fermi level at the surface within several kT, or even into, the silicon conduction and valence bands during stress, so that Boltzmann statistics cannot be used. Thus, a correct solution for ΔV_T must, out of necessity, be performed numerically.

The model discussed in this and the previous chapter has therefore been implemented in a software simulation program, which is described in Chapter 7. The simulation program is valid for arbitrary trap distributions, bias conditions and temperatures.

While the exact problem does not lend itself to simple closed form solutions of general validity, it is possible to obtain approximate solutions under suitably restricted conditions of bias and time. In this section we make several simplifying assumptions in order to obtain some insight into the functional form of the voltage and time dependence of the threshold voltage shifts. The simplified problem which we consider is representative of the conditions under which most of our experiments are carried out.

We assume an NMOS transistor in the linear region, initially biased in inversion or strong inversion, so that the Fermi level is in the upper half of the bandgap, close

to the conduction band. During stress, the Fermi level can be driven up into the conduction band by several hundredths of an electron volt, but for this approximate discussion we can assume it to be pinned at $E_F \approx E_c$.

The following assumptions are also made:

1. Uniform oxide trap distribution in energy and space:

$$D_{ot}(E,x) = D_{ot} \quad (\mathrm{cm}^{-3}\ \mathrm{eV}^{-1}) \tag{4.50}$$

2. Rectangular tunneling barrier with a time constant independent of energy and electric field:

$$\tau_{ot}(x) = \tau_{oc} e^{x/x_o} \tag{4.51}$$
$$\tau_{ot}{}^{o}(x) = \tau_{oe} e^{x/x_o} \tag{4.52}$$

where $x_o = \hbar/2\sqrt{2M^*(\phi_c - E)} \approx \hbar/2\sqrt{2M^*\phi_c}$. This is valid since we are only interested in electrons near the conduction band edge $E \ll \phi_c$. We allow for the possibility that the prefactors during stress (τ_{oc}) and after stress (τ_{oe}) may differ.

3. Step approximation to the Fermi function:

$$f(E) = \begin{cases} 1 & E < E_F \\ 0 & E > E_F \end{cases} \tag{4.53}$$

4. Maximum trapping depth $x_m \approx x_o \ln(t_s/\tau_{oc}) \ll t_{ox}$

The information that can be obtained from such a simplified model is limited, but shall prove useful for qualitatively understanding the nature of the voltage and time dependence. Such a model will not provide complete information on temperature dependence or on the effects of a nonuniform distribution of traps, for which the exact model must be employed. The temperature dependence will be examined separately in the next section.

The third assumption implies that

$$f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x) = \begin{cases} -1 & E_F - q\mathcal{E}x < E_t < E_{F0} - q\mathcal{E}_o x \\ 1 & E_{F0} - q\mathcal{E}_o x < E_t < E_F - q\mathcal{E}x \\ 0 & \text{otherwise} \end{cases} \tag{4.54}$$

The shaded regions in Fig. 4.8 indicate those trap energies for which $\Delta f = f - f_0 = 1$ and which, therefore, contribute to the threshold voltage shift. The case $\Delta f = -1$ will not occur for positive stress voltage pulses $V_p > 0$ since $E_F > E_{F0}$ and $\mathcal{E} < \mathcal{E}_o$. Trap energies for which $\Delta f = 0$ are either always empty or always filled and therefore

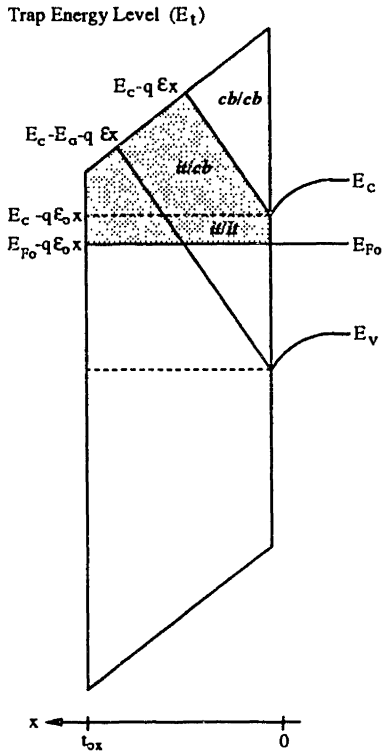


Figure 4.8: Band diagram for an NMOS transistor pulsed with a positive gate voltage.

do not contribute to the threshold voltage shift. Note that, under the assumption of a step Fermi function, $\Delta f = 0$ for the cb/cb component, a problem which is addressed below.

While the general formulation of the problem, as given by Eqs. 4.41–4.43 and 4.47–4.49, is independent of whether oxide traps between the silicon band edges charge by the two-step or lattice relaxation processes, an explicit solution of these equations requires the specific form of the relaxation transients R_i. Since the dynamics of the two-step and lattice relaxation mechanisms differ. each of these is considered separately in the next two sections.

4.4.1 Two-Step Process

In this section, approximate solutions are derived assuming that oxide traps above the silicon conduction band communicate by elastic tunneling with conduction band states, while those below capture and emit by the two-step process. The form of the solutions for the three dominant components is discussed below.

Interface State/Interface State Component (it/it)

Under the assumed bias conditions, only tunneling from fast surface states near the conduction band is important, for which the thermal time constant is much less than the tunneling time constant, $\tau_{it} \ll \tau_{ot/it}$. In this regime, the single time constant model (approximation 1) of Eqs. 3.77 and 3.78 is valid. Equation 4.43 for the it/it component, can be written, under these assumptions, as

$$\Delta V_T(t_s,t_r) = \frac{qD_{ot}t_{ox}}{\epsilon_{ox}} \int_0^{x_1} \int_{E_c-E_G-q\mathcal{E}x}^{E_F-q\mathcal{E}_o x} (f-f_0)[1-e^{-t_s/\tau_{ot}(x)}]e^{-t_r/\tau_{ot}{}^o(x)} dE_t dx \quad (4.55)$$

where we have used the fact that $E_F \approx E_c$ during stress. The integration limits are shown in Fig. 4.8. With a step Fermi function, $f - f_o = 0$ for $E_t < E_{Fo} - q\mathcal{E}_o x$, so that the it/it component is nonzero close to the interface only over the range $E_{F0} - q\mathcal{E}_o x < E_t < E_F - q\mathcal{E}_o x$. Thus,

$$\Delta V_T(t_s,t_r) = \frac{qD_{ot}t_{ox}}{\epsilon_{ox}} \int_0^{x_1} \int_{E_{F0}-q\mathcal{E}_o x}^{E_F-q\mathcal{E}_o x} [1-e^{-t_s/\tau_{ot}(x)}]e^{-t_r/\tau_{ot}{}^o(x)} dE_t dx \quad (4.56)$$

Since the time constants are assumed to be independent of energy, the energy integration can be carried out to give

$$\begin{aligned} \Delta V_T(t_s,t_r) &= \frac{qD_{ot}t_{ox}}{\epsilon_{ox}}(E_F - E_{F0}) \int_0^{x_1} [1-e^{-t_s/\tau_{ot}(x)}]e^{-t_r/\tau_{ot}{}^o(x)} dx \\ &= \frac{qD_{ot}t_{ox}}{\epsilon_{ox}}(E_F - E_{F0}) \int_0^{x_1} g(x,t_s)[1-g^o(x,t_r)]dx \end{aligned} \quad (4.57)$$

where, following Heiman and Warfield [58], we define

$$g(x,t) \equiv 1 - \exp[-(t/\tau_{oc})\exp(-x/x_o)] \approx \begin{cases} 1 & x < x_o \ln(t/\tau_{oc}) \\ 0 & x > x_o \ln(t/\tau_{oc}) \end{cases} \quad (4.58)$$

$$1 - g(x,t) = \exp[-(t/\tau_{oc})\exp(-x/x_o)] \approx \begin{cases} 0 & x < x_o \ln(t/\tau_{oc}) \\ 1 & x > x_o \ln(t/\tau_{oc}) \end{cases} \quad (4.59)$$

with $g^o(x,t)$ defined in an analogous manner, with τ_{oe} replacing τ_{oc}.

The function $g(x,t_s)$, which is plotted in Fig.4.9(a), resembles a step function in the spatial variable x, falling off abruptly at $x = x_o \ln(t_s/\tau_{oc})$. The logarithmic dependence of the depth to which traps fill on stress time is readily apparent. Traps fill to a depth of approximately $15\mathring{A}$ after 10ms of stress. Figure 4.9(b) plots the integrand of Eq. 4.57, $g(x,t_s)[1-g^o(x,t_r)]$, for a stress time of $t_s = 10ms$. The decay of the initially filled trap distribution For $t_r \ll t_s$ the function appears rectangular

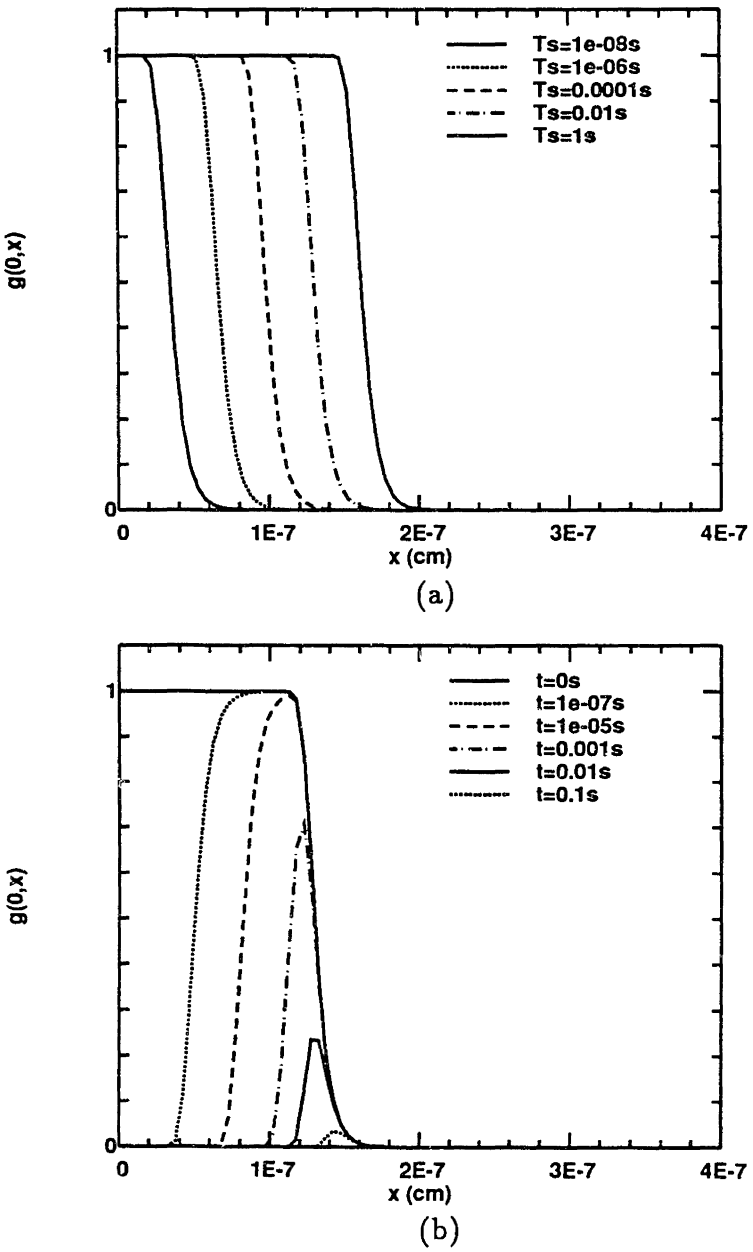


Figure 4.9: Spatial variation of the filled oxide trap density with time (a) $g(x, t_s)$ during stress, (b) $g(x, t_s)[1 - g^o(x, t_r)]$ during emission, after $t_s = 10$ms.

and the traps can be approximated as being filled for $x_e < x < x_f$, where x_f and x_e are the depths to which traps fill during stress and empty during emission, given by

$$x_f = x_o \ln\left(\frac{t_s}{\tau_{oc}}\right) \tag{4.60}$$

$$x_e = x_o \ln\left(\frac{t_r}{\tau_{oe}}\right) \tag{4.61}$$

With these approximations, the integral in Eq. 4.57 is approximately unity for $x_o \ln(t_r/\tau_{oe}) < x < x_o \ln(t_s/\tau_{oc})$ so that

$$\Delta V_T(t_s, t_r) = \frac{qD_{ot}x_o}{C_{ox}}(E_F - E_{F0})\left[\ln\left(\frac{t_s}{\tau_{oc}}\right) - \ln\left(\frac{t_r}{\tau_{oe}}\right)\right] \tag{4.62}$$

where $C_{ox} = \epsilon_{ox}/t_{ox}$ is the oxide capacitance per unit area. Equation 4.62 simply states that the threshold voltage shift is the total trapped charge, $Q_{ot} = qD_{ot}\Delta E \Delta x$ divided by the oxide capacitance, where $\Delta E = E_F - E_{F0}$ is the energy range of traps swept out by the change in Fermi level and $\Delta x = x_f(t_s) - x_e(t_r)$ is the difference between the depth to which traps fill in time t_s and the depth to which they empty in time t_r.

It is clear from Fig. 4.9(b) that approximating the filled trap density by a rectangular distribution with sharp cutoffs, is only valid for $t_r \ll t_s$. This is almost always the case during stress, when t_r is only due to circuit delays. Thus, the threshold voltage shift should be well modeled as a logarithmic function of stress time. The relaxation transient is also approximately logarithmic until t_r becomes comparable to t_s. Due to the tails of the exponential distributions in Eqs. 4.58 and 4.59, some traps with $x > x_f(t_s)$ will be filled and these traps will empty when $t_r > t_s$. Figure 4.9(b) shows that, after 10ms of stress followed by 10ms of relaxation, the only remaining trapped charge is a small fraction of filled traps resembling a pulse at at about 12$\mathring{A}$ from the interface. The long tail decays which are observed in experiments persist for hundreds of milliseconds and are due to the relaxation of this residual trapped charge from the deepest traps which are partially filled during stress. Since the approximations, Eqs. 4.58 and 4.59 clearly break down in this regime, the results obtained in this section must be regarded with suspicion and the exact model must be used for an accurate interpretation of experimental long-tails. In order to get some idea of the error in the approximation used above, we repeat the x integration more carefully.

We require the integral

$$\int_0^{x_1} \exp\left[-\frac{t}{\tau_o}\exp(-x/x_o)\right] \tag{4.63}$$

Changing the variable of integration from x to $y = \frac{t}{\tau_o}\exp(-x/x_o)$, the integral becomes

$$\int_0^{x_1} \exp\left[-\frac{t}{\tau_o}\exp(-x/x_o)\right] = \int_{t/\tau_1}^{t/\tau_o} \frac{e^{-y}}{y} dy$$

$$= x_o\left[E_1\left(\frac{t}{\tau_1}\right) - E_1\left(\frac{t}{\tau_o}\right)\right] \tag{4.64}$$

where $\tau_1 = \tau_o \exp(x/x_1)$ and $E_1(z)$ is the exponential integral function, [158]

$$E_1(z) = \int_z^\infty \frac{e^{-x}}{x} dx \tag{4.65}$$

which has the properties that $\lim_{z\to\infty} E_1(z) = 0$ and has the series expansion

$$\begin{aligned} E_1(z) &= -\gamma - \ln z - \sum_{n=1}^{\infty} \frac{(-1)^n z^n}{nn!} \\ &\approx -\gamma - \ln z \qquad \text{for small } z \end{aligned} \tag{4.66}$$

where $\gamma = 0.5772$ is Euler's constant.

With these considerations, Eq. 4.57 can be written

$$\Delta V_T(t_s, t_r) = \frac{qD_{ot}x_o}{C_{ox}}(E_F - E_{F0})\left\{E_1\left(\frac{t_r}{\tau_{1e}}\right) - E_1\left(\frac{t_r}{\tau_{oe}}\right) + E_1\left(\frac{t_s}{\tau_{1c}} + \frac{t_r}{\tau_{1e}}\right) - E_1\left(\frac{t_s}{\tau_{oc}} + \frac{t_r}{\tau_{oe}}\right)\right\} \tag{4.67}$$

where $\tau_{1e} = \tau_{oe}\exp(x_1/x_o)$ and $\tau_{1c} = \tau_{oc}\exp(x_1/x_o)$ It is always the case in our measurements that stress times and circuit delays are much greater than the smallest tunneling time constant, which may be on the order of tens of picoseconds, thus $t_s \gg \tau_{oc}$ and $t_r \gg \tau_{oe}$ Thus, the second and third bracketed terms in Eq. 4.67 can be neglected. Since x_1, defined by Eq. 4.38 is generally about $50\mathring{A}$ under our measurement conditions, the time constants τ_{1c} and τ_{1e} are very large relative to time scales of the measurement, so that Eq. 4.66 can be used to rewrite Eq 4.67 as

$$\Delta V_{Tit/it}(t_s, t_r) = \frac{qD_{ot}x_o}{C_{ox}}(E_F - E_{F0})\ln\left(1 + \frac{\tau_{oe}t_s}{\tau_{oc}t_r}\right) \tag{4.68}$$

For $\tau_{oc} \approx \tau_{oe}$, when $t_r \ll t_s$, Eq. 4.68 reduces to Eq. 4.62. When $t_r \gg t_s$ then, retaining only the first term in the Taylor series expansion

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \ldots \tag{4.69}$$

we obtain the asymptotic expressions,

$$\Delta V_{Tit/it}(t_s, t_r) = \begin{cases} \frac{qD_{ot}x_o}{C_{ox}}(E_F - E_{F0})\left[\ln\left(\frac{t_s}{\tau_{oc}}\right) - \ln\left(\frac{t_r}{\tau_{oe}}\right)\right] & \tau_{oe} < t_r \ll t_s \\ \frac{qD_{ot}x_o}{C_{ox}}(E_F - E_{F0})\frac{\tau_{oe}t_s}{\tau_{oc}t_r} & t_r \gg t_s \end{cases} \tag{4.70}$$

This more exact calculation indicates that there will be deviations from the simple logarithmic time dependence when the relaxation time t_r exceeds the stress time t_s,

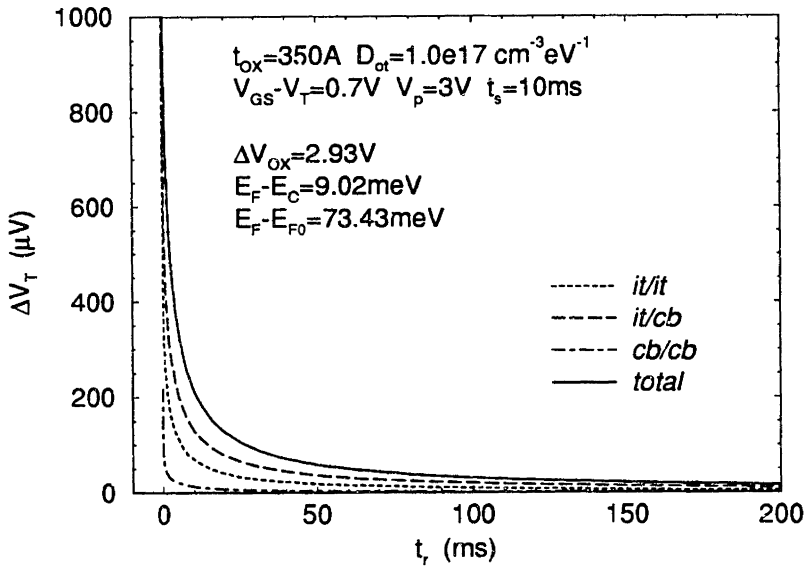


Figure 4.10: Simulated relaxation transient showing components it/it (Eq. 4.68), it/cb (Eq. 4.77) and cb/cb (Eq. 4.81).

and that this long time behavior will approach $1/t_r$. The $1/t$ dependence is not predicted by other derivations of the time dependence of the emission of trapped charge [159, 160, 161] which predict a purely logarithmic decay. The reason for the discrepency is that previous calculations assume that all traps are initially filled rather than including the filling transient as was done above.

Figure 4.10 plots Eq. 4.68 for an NMOS transistor with $t_{ox} = 350\mathring{A}$ stressed with $V_p = 3$V for a period of $t_s = 10$ms. On a linear-log scale, the emission transient is a straight line for $t_r \ll t_s$ as shown in Fig. 4.11(a) and indicated by the logarithmic asymptote of Eq. 4.70. Some curvature becomes evident at longer times. Figure 4.11(b) shows the same simulation plotted on a log-log scale. In this figure the function plots as a straight line with unity slope for $t_r \gg t_s$ indicating a $1/t$ decay as predicted by Eq. 4.70. The region of interest in our measurements occurs around the knee of the emission curve ($t_r \approx t_s$) where neither of the asymptotic expressions is valid and the transient exhibits substantial curvature.

The gate voltage dependence of the it/it component is determined by the change in Fermi level at the interface $\Delta E_F = E_F - E_{F0} = q\Delta\psi_s$, where $\Delta\psi_s$ is the change in surface potential that occurs during stress. The surface potential cannot be written explicitly as a function of gate voltage $V_G' = V_G - V_{FB}$ in a MOS structure. Nevertheless, some qualitative conclusions can be drawn from the plot of surface potential versus gate voltage shown in Fig. 4.12. This graph was obtained by a numerical solution of Eqs. E.7 and E.8. Note that the width along the vertical axis between the saturating tails of the characteristic is wider than the silicon bandgap (1.12V),

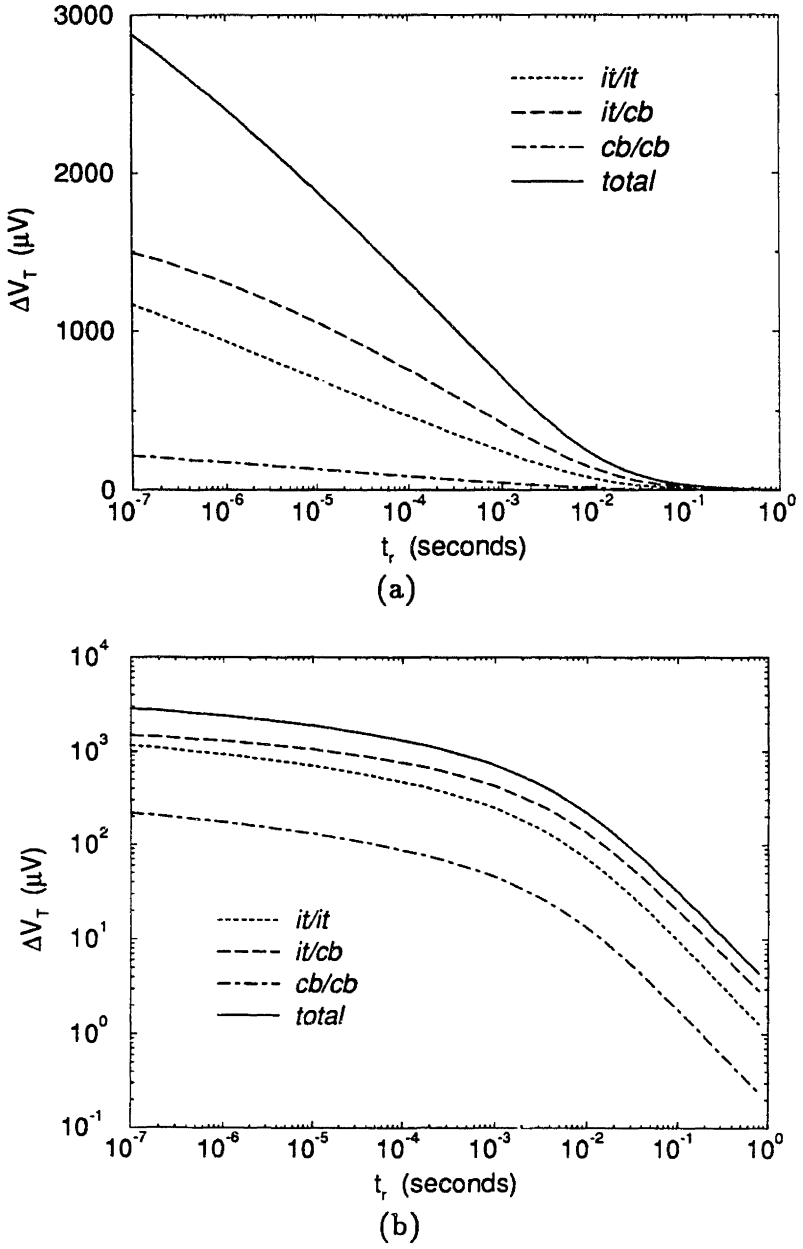


Figure 4.11: Relaxation transient of Fig. 4.10 plotted on (a) log-linear axes (b) log-log axes.

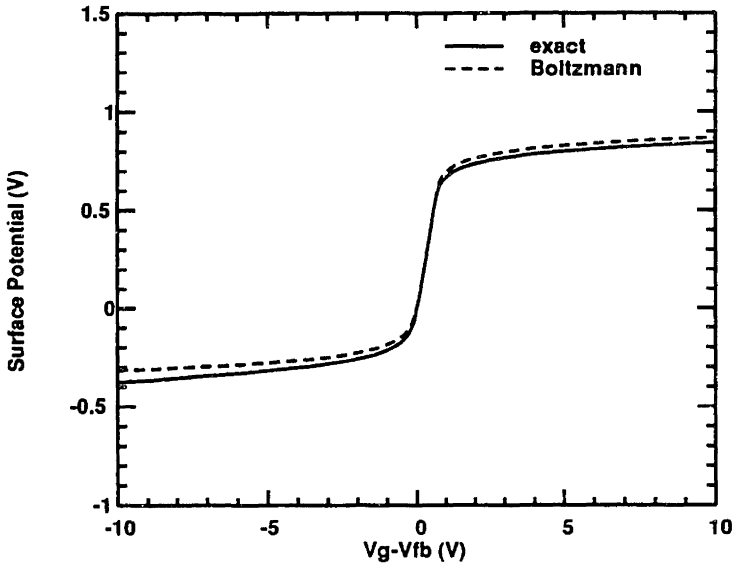


Figure 4.12: Exact and Boltzmann approximations for the surface potential versus gate voltage for a 250$\AA$ gate oxide with substrate doping of $N_A = 1.0 \times 10^{15}\text{cm}^{-3}$.

indicating that the Fermi level is driven into the conduction and valence bands during positive and negative gate voltage stresses. This is why Boltzmann approximations fail and a numerical solution is necessary.

The surface potential changes rapidly with $V_G{}'$ in depletion and weak inversion and hence the it/it component will vary rapidly with voltage as the Fermi level sweeps through the bandgap. In strong inversion, $U_s > 2U_F$, the surface potential changes very slowly with gate voltage. In this region, the Fermi level tends to become pinned which tends to suppress the it/it component.

Since we are interested in large signal gate voltage excursions, there are no generally valid approximations that can be used for the gate voltage – surface potential relationship. Even for the common case of inversion region operation, with positive gate voltage pulses as discussed here, the Fermi level is generally within a few kT below or above the conduction band edge before and after stress, so that we are operating at the knee of the curve in Fig. 4.12 in the non-Boltzmann regime, where most approximate expressions break down. Therefore no attempt is made to write the explicit voltage dependence of the threshold shift components. Figure 4.14 plots the voltage dependence of Eq. 4.68 for $V_p > 0$ using the numerical solution to Eqs. E.7 and E.8.

The results for the it/it component for a uniform trap distribution, under the assumptions stated above can be summarized as follows:

1. The amplitude of the threshold voltage shift is directly proportional to the density of oxide traps D_{ot}.

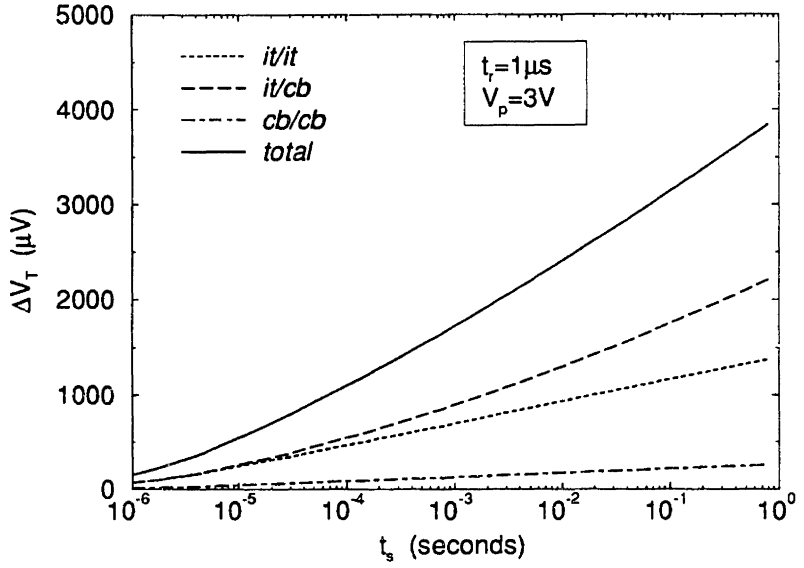


Figure 4.13: Components of threshold voltage shift versus stress time t_s.

2. The amplitude of the threshold voltage shift increases logarithmically with stress time t_s.

3. The relaxation of the threshold voltage shift with time t_r after stress is approximately logarithmic for $t_r \ll t_s$ and approaches $1/t_r$ for $t_r \gg t_s$.

4. The amplitude of the threshold voltage shift is proportional to the change in the Fermi level $\Delta E_F = E_F - E_{F0}$ at the interface during stress.

Interface State/Conduction Band Component (it/cb)

The it/cb component [Eq. 4.42], under the aforementioned assumptions, becomes

$$\Delta V_T(t_s, t_r) = \frac{qD_{ot}t_{ox}}{\epsilon_{ox}} \int_0^{x_1} \int_{E_c - q\mathcal{E}_o x}^{E_c - q\mathcal{E}x} (f - f_0)[1 - e^{-t_s/\tau_{ot}(x)}]e^{-t_r/\tau_{ot}^{\circ}(x)} dE_t dx \qquad (4.71)$$

From Fig. 4.8, $f - f_0 = 1$ over the entire range of the energy integration so that it can be carried out directly to give

$$\Delta V_T(t_s, t_r) = \frac{-q^2 D_{ot}t_{ox}}{\epsilon_{ox}}(\mathcal{E} - \mathcal{E}_o) \int_0^{x_1} x[1 - e^{-t_s/\tau_{ot}}]e^{-t_r/\tau_{ot}^{\circ}} dx \qquad (4.72)$$

Using the approximations of Eqs. 4.58 and 4.59 and making the reasonable assumption that x_1 is much larger than the maximum tunneling depth $x_m = x_o \ln(t_s/\tau_{oc})$,

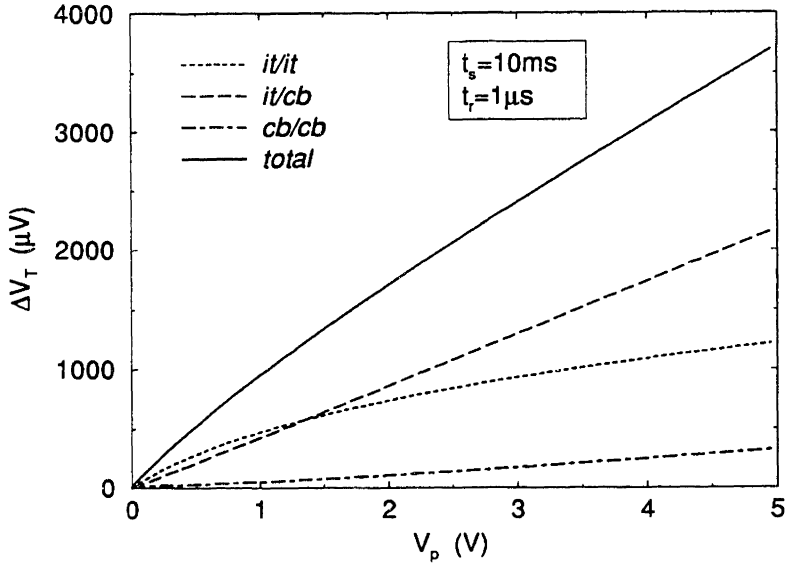


Figure 4.14: Components of threshold voltage shift versus stress voltage V_p.

Eq. 4.72 can evaluated as

$$\begin{aligned}\Delta V_{T\,it/cb}(t_s,t_r) &= \frac{q^2 D_{ot}}{\epsilon_{ox}}\Delta V_{ox}\int_{x_o\ln(t_s/\tau_{oc})}^{x_o\ln(t_r/\tau_{oe})} x\,dx \\ &= \frac{q^2 D_{ot} x_o^2}{2\epsilon_{ox}}\Delta V_{ox}[\ln^2(t_s/\tau_{oc}) - \ln^2(t_r/\tau_{oc})] \end{aligned} \tag{4.73}$$

where the voltage drop across the oxide is defined in the opposite sense as the electric field, $V_{ox} = V(t_{ox}) - V(0) = -\mathcal{E}_{ox}t_{ox}$. Equation 4.73 again has a simple interpretation. Referring to Fig. 4.8, the range of traps which fall below the conduction band during stress and above it after stress are those which fall into a triangle of length $x_f(t_s)$ in space, and height along the energy axis of $\Delta E = -q\Delta\mathcal{E}x_f(t_s)$. The trapped charge is just the area of this triangle $Q_{ot} = qD_{ot}\frac{1}{2}[q\Delta\mathcal{E}x_f(t_s)][x_f(t_s)]$ The traps which empty after stress is given by a similar expression with $x_f(t_s)$ replaced by $x_e(t_r)$. The threshold voltage shift at any time is just the difference between these two divided by C_{ox}, as in Eq. 4.73. The $\ln^2$(time) dependence is a consequence of the fact that the spatial depth to which traps fill goes as ln(time) and the range of energies which are filled as a result of oxide band bending is also proportional to x and hence, ln(time).

We now undertake a more exact evaluation of the integral (4.72) in order to determine the behavior of the threshold voltage relaxation at long times $t_r \gg t_s$, for which the approximations, Eqs. 4.58 and 4.59, are not valid. First note that

$$-\frac{d}{dt}\int_0^{x_1} xe^{-t/\tau}dx = \int_0^{x_1} xe^{-t/\tau}\frac{dx}{\tau} \tag{4.74}$$

The right hand side can easily be evaluated, using the substitution $u = t/\tau$,

$$\begin{aligned}\int_0^{x_1} xe^{-t/\tau}\frac{dx}{\tau} &= \frac{x_o^2}{t}\int_{t/\tau_1}^{t/\tau_o}\left[\ln\left(\frac{t}{\tau_o}\right) - \ln(u)\right]e^{-u}du \\ &= \frac{x_o^2}{t}\left[\Gamma(1)\ln\left(\frac{t}{\tau_o}\right) - \Gamma'(1)\right] \end{aligned} \tag{4.75}$$

where $\tau_1 = \tau_o \exp(x_1/x_o)$ and for the time scales of interest during the relaxation, $\tau_o \ll t \ll \tau_1$, so that the lower and upper integration limits can be replaced by zero and infinity, respectively. From Eq. 4.74, and using $\Gamma(1) = 1$ and $\Gamma'(1) = -\gamma$,

$$\int_0^{x_1} xe^{-t/\tau}dx = -x_o^2\left[\frac{1}{2}\ln^2\left(\frac{t}{\tau_o}\right) + \gamma\ln\left(\frac{t}{\tau_o}\right)\right] \tag{4.76}$$

Using Eqs. 4.72 and 4.76, the it/cb contribution to the threshold voltage shift is

$$\begin{aligned}\Delta V_{Tit/cb}(t_s,t_r) &= \frac{q^2 D_{ot}x_o^2}{\epsilon_{ox}}\Delta V_{ox}\left\{\frac{1}{2}\left[\ln^2\left(\frac{t_s}{\tau_{oc}}+\frac{t_r}{\tau_{oe}}\right) - \ln^2\left(\frac{t_r}{\tau_{oe}}\right)\right]\right. \\ &\qquad \left. +\gamma\left[\ln\left(\frac{t_s}{\tau_{oc}}+\frac{t_r}{\tau_{oe}}\right) - \ln\left(\frac{t_r}{\tau_{oe}}\right)\right]\right\} \\ &= \frac{q^2 D_{ot}x_o^2}{\epsilon_{ox}}\Delta V_{ox}\ln\left(1+\frac{\tau_{oe}t_s}{\tau_{oc}t_r}\right)\left[\gamma + \frac{1}{2}\ln\left(1+\frac{\tau_{oe}t_s}{\tau_{oc}t_r}\right) + \ln\left(\frac{t_r}{\tau_{oe}}\right)\right]\end{aligned} \tag{4.77}$$

For $t_r \ll t_s$, keeping only the quadratic term, Eq. 4.77 reduces to the previous result, Eq. 4.72. For $t_r \gg t_s$, Eq. 4.77 can be approximated using the first term of the Taylor series expansion, Eq. 4.69. Thus, the asymptotic expressions for the it/cb component are

$$\Delta V_{Tit/cb}(t_s,t_r) = \begin{cases} \frac{q^2 D_{ot}x_o^2}{2\epsilon_{ox}}\Delta V_{ox}\left[\ln^2\left(\frac{t_s}{\tau_{oc}}\right) - \ln^2\left(\frac{t_r}{\tau_{oe}}\right)\right] & \tau_{oe} < t_r \ll t_s \\ \frac{q^2 D_{ot}x_o^2}{\epsilon_{ox}}\Delta V_{ox}\frac{\tau_{oe}t_s}{\tau_{oc}t_r}\left[\gamma + \ln\left(\frac{t_r}{\tau_{oe}}\right) + \frac{1}{2}\left(\frac{\tau_{oe}t_s}{\tau_{oc}t_r}\right)\right] & t_r \gg t_s \end{cases} \tag{4.78}$$

At long times, we can therefore expect to observe deviations from the simple $\ln^2$ time dependence of Eq. 4.72, including components which go as $1/t$ and $1/t^2$. Equation 4.77 is plotted in Figs. 4.10–4.14.

In summary, the it/cb component has the following properties:

1. The threshold voltage shift is proportional to the oxide trap density D_{ot}.
2. The threshold voltage shift as a function of stress time t_s goes as $\ln^2(t_s)$, which appears concave upward on a log time scale [Fig. 4.13].
3. The relaxation of the threshold voltage shift as a function of time t_r after removal of the stress goes as $\ln^2(t_r)$ for $t_r \ll t_s$. For $t_r \gg t_s$, the relaxation tends to $1/t_r$ as shown in Fig. 4.11.

4. The threshold voltage shift is proportional to the change in voltage across the oxide $\Delta V_{ox} = \Delta V_G - \Delta\psi_s$ during stress. In depletion and weak inversion when ψ_s changes rapidly with gate voltage, as shown in Fig. 4.12, the it/cb component is relatively weak. In strong inversion, the surface potential is nearly constant so that most of the applied gate voltage is dropped across the oxide $\Delta V_{ox} \approx \Delta V_G$. In this regime the it/cb contribution to the threshold voltage shift is directly proportional to the amplitude of the applied voltage pulse $V_p = \Delta V_G$ as shown in Fig. 4.14.

Conduction Band/Conduction Band Component (cb/cb)

The cb/cb component results from the tunneling of electrons from filled conduction band states into oxide traps. The approximation, Eq. 4.53, used in the previous two examples cannot be applied to this case because, the step approximation implies that there are no electrons in conduction band when $E_F < E_c$. It is therefore necessary to use the full Fermi function rather than a step approximation.

The threshold voltage shift is given by Eq. 4.41 as

$$\Delta V_T(t_s, t_r) = \frac{qD_{ot}t_{ox}}{\epsilon_{ox}} \int_0^{x_2} \int_{E_c - q\mathcal{E}x}^{E_c+\phi_c} [f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x)] \times [1 - e^{-t_s/\tau_{ot}(x)}] e^{-t_r/\tau_{ot}^o(x)} dE_t dx$$

The upper limit of integration can be replaced by infinity since the Fermi functions in the integrand are vanishingly small at $E_t = E_c + \phi_c$. Since the time constant is assumed independent of energy, the energy integral then can be carried out, using Eqs. 4.18 and 4.19, to give

$$\Delta V_T(t_s, t_r) = \frac{qD_{ot}t_{ox}}{\epsilon_{ox}} \int_0^{x_2} N(x)[1 - e^{-t_s/\tau_{ot}(x)}] e^{-t_r/\tau_{ot}^o(x)} dx \tag{4.79}$$

where

$$\begin{aligned} N(x) &\equiv \int_{E_c - q\mathcal{E}x}^{\infty} [f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x)] dE_t \\ &= kT \ln \left\{ \frac{1 + \exp(E_F - E_c)/kT}{1 + \exp(E_{F0} - E_c + q\Delta\mathcal{E}x)/kT} \right\} \end{aligned} \tag{4.80}$$

where $\Delta\mathcal{E} = \mathcal{E} - \mathcal{E}_o$ is the change in electric field across the oxide during stress. The initial bias is usually such that $E_{F0} - E_c < 3kT$ and for positive pulses, $q\Delta\mathcal{E}x < 0$ so that the exponential term in the denominator of $N(x)$ is normally much less than one, and can be neglected. $N(x)$ can then be taken outside the integral and the x integration carried out in a manner identical to the it/it case above, to give

$$\Delta V_{Tcb/cb}(t_s, t_r) \approx \frac{qD_{ot}x_o}{C_{ox}} kT \ln \left\{ 1 + \exp\left[\frac{E_F - E_c}{kT} \right] \right\} \ln \left(1 + \frac{\tau_{oe}t_s}{\tau_{oc}t_r} \right) \tag{4.81}$$

The form of this result is identical to that of the *it/it* component, Eq. 4.68 except that the energy range is given by

$$\Delta E = kT \ln \left\{ 1 + \exp \left[\frac{E_F - E_c}{kT} \right] \right\} \tag{4.82}$$

Equation 4.81 is plotted in Figs. 4.10–4.14. For the stated bias conditions and trap density, the *cb/cb* component is smaller than the other two, however this may not always be the case.

The following conclusions can be drawn from Eq. 4.81 for the *cb/cb* component:

1. The threshold voltage shift is proportional to the oxide trap density D_{ot}.
2. The threshold voltage shift as a function of stress time goes as $\ln(t_s)$ [Fig. 4.13].
3. The relaxation of the threshold volage shift with time following stress goes as $\ln(t_r)$ for $\tau_o < t_r < t_s$ and approaches $1/t_r$ for $t_r > t_s$ [Fig. 4.11].
4. For $E_F - E_c <$ several kT the exponential factor in ΔE is small and the energy range of the traps which contribute to the threshold voltage shift resembles an impulse of width kT around the conduction band edge. For $E_F - E_c >$ several kT, the surface of the silicon becomes degenerate and $\Delta E \approx E_F - E_c$. In this limit, the result resembles the case of a metal where tunneling occurs in the energy range between the bottom of the conduction band the Fermi level.

4.4.2 Lattice Relaxation

The approximate strong inversion model of Sec. 4.4.1 requires only a minor modification in order to incorporate lattice relaxation. Only the prefactors of the time constants for lattice relaxation and the two-step process differ [Eqs. 3.34, 3.44, 3.45, 3.99, 3.101]. In strong inversion, $n_s \gg n_1, p_1, p_s$ and the total inversion layer charge Q_N (coul-cm^{-2}) is just $C_{ox}(V_{GS} - V_T)$. As a first order approximation we can write

$$n_s = \frac{C_{ox}(V_{GS} - V_T)}{q\bar{x}_{inv}} \tag{4.83}$$

where $\bar{x}_{inv}$ (cm) can be interpreted as the mean thickness of the inversion layer. Oxide traps below the conduction band are assumed to communicate with it by the tunneling/LRME mechanism while those above require no thermal assistance since conduction band states at the same energy are accessible by elastic tunneling. Using Eq. 4.83 in Eq. 3.99, the time constant prefactors in Eqs. 4.51 and 4.52 become

$$\tau_{oc} = \frac{e^{\Delta E_{Bc}/kT}}{\alpha(V_{GS} - V_T)} \tag{4.84}$$

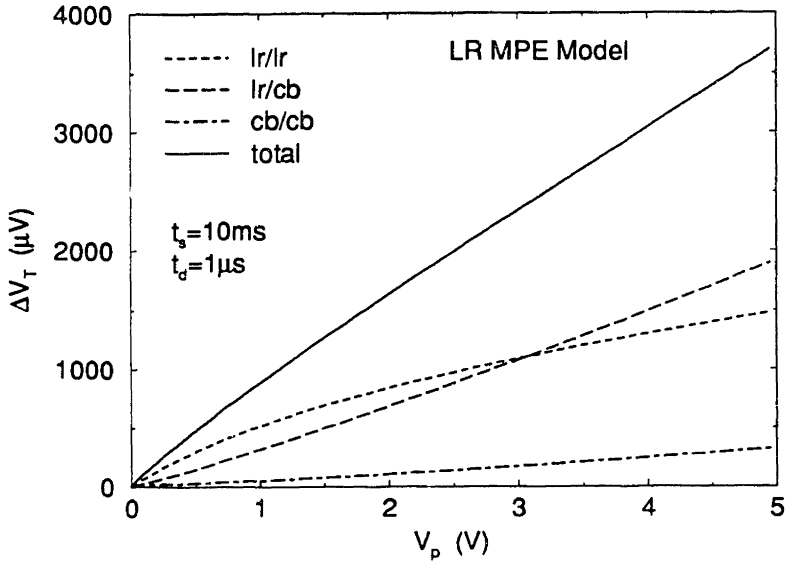


Figure 4.15: Components of threshold voltage shift versus stress voltage V_p using lattice relaxation multiphonon emission model.

for the lr/lr and lr/cb components, and

$$\tau_{oe} = \frac{e^{\Delta E_{Be}/kT}}{\alpha(V_{GS}^{o} - V_T)} \tag{4.85}$$

in the lr/lr component, where V_{GS} and V_{GS}^{o} are the gate-source voltage during and before/after stress, respectively, and

$$\alpha = C_{ox}\sigma_n\overline{v}/q\overline{x}_{inv} \tag{4.86}$$

is a constant with units $s^{-1}V^{-1}$. The prefactors in the remaining cases remain unchanged. With these substitutions the results of Sec. 4.4 are directly applicable to lattice relaxation, since the lattice relaxation prefactors contain no dependence on energy or space. The only feature of the prefactors 4.84 and 4.85 which differs from the two-step case is their dependence on voltage. Thus, at a given gate bias, the functional dependence of ΔV_T on stress and relaxation time, as shown in Figs. 4.10–4.13, remains valid for lattice relaxation. Only the dependence on stress voltage needs to be examined.

Figure 4.15 plots the threshold voltage components, Eqs. 4.68, 4.77 and 4.81, with the substitutions specified by Eqs. 4.84 and 4.85. The barrier height ΔE_B is taken to be zero, an assumption which can, at most, affect the amplitude, but not the voltage dependence. Values used for the other model parameters were $\alpha = 1.0 \times 10^8 s^{-1}$ and $\tau_{oc} = \tau_{oe} = 1.0 \times 10^{-8}$s. It can be seen that the lr/lr and lr/cb components change

slightly faster with voltage than do the it/it and it/cb components in the two-step model [Fig. 4.14], due to the voltage dependence in the time constant prefactor. However, the factor of five increase in the prefactor over the plotted voltage range results in a negligible change in the total threshold voltage shift. We can therefore conclude from this approximate model that the essential features of the threshold voltage relaxations are independent of which of these two mechanisms actually dominates. Due to the activated term ΔE_B in the lattice relaxation time constant, the temperature dependence of the two mechanisms will differ, an issue which is addressed in Chap. 7.

4.5 Temperature Dependence

The tunneling probability, as determined by the WKB factors in Secs. 3.4–3.6, is nearly independent of temperature. However, the overall rate at which oxide traps charge and discharge may depend on temperature through the following:

1. The correlated mobility term in Eq. 4.13.
2. The distribution of electrons available for tunneling, as determined by the fermi function and the density of states in the semiconductor.
3. The thermally-activated time constant for interface state-assisted tunneling [Eq. 3.102].
4. The thermally-activated time constant for lattice relaxation multiphonon emission [Eq. 3.99].

In addition, there are several additional sources of temperature dependence which are included in the numerical simulation model, as described in Sec. 7.2.4, but are not significant in the extrinsic temperature range of interest. These include the temperature dependences of the energy gap, the effective masses and the ionized dopant concentrations. Temperature dependence can affect either the amplitude or the time constant of the equivalent gate voltage shift. The contributions of each of the primary temperature-dependent factors are considered below, assuming for the purposes of discussion, an NMOS transistor.

4.5.1 Correlated Mobility Fluctuations

The temperature dependence of the correlated mobility term has already been discussed in the previous section. The effect of mobility fluctuations depends on whether the traps are neutral or charged when filled. Acceptor-like traps in an NMOS transistor will be filled during a positive (inverting) pulse, becoming negatively charged and increasing coulombic scattering of inversion layer electrons when the pulse is removed. If the mobility term in Eq. 4.15 is large compared with the threshold voltage term,

as may be the case in NMOS transistors [118], correlated mobility fluctuations can lead to a decrease in the amplitude of ΔV_{GS} with temperature which goes as $T^{-3/2}$. For negative pulses, acceptor-like traps would be emptied, becoming neutral, which would tend to increase the mobility and decrease the equivalent gate voltage shift. The gate voltage shift would then increase with temperature for negative pulses.

It has been shown, however, that traps in the upper half of the bandgap tend to be acceptor-like, while those in the lower half are donor-like [157]. Donor-like traps would be emptied during negative pulses becoming positively charged. The resulting decrease in mobility would increase the magnitude of the equivalent gate voltage shift, an effect which would decrease with increasing temperature.

In PMOS transistors, the effect of correlated mobility fluctuations is generally negligible [118]. The contribution of mobility fluctuations to the overall temperature dependence is complicated by uncertainties concerning the relative importance of mobility and number fluctuations and specifically, the value of S_o in Eq. 4.15.

4.5.2 Carrier Distributions in the Semiconductor

The tunneling rate at a given energy depends on the concentration of free charge carriers available to tunnel at that energy. Previous researchers have treated the temperature dependence of the capture and emission of charge from oxide traps in the context of the Heiman and Warfield model by writing the electron concentration at energy E above the conduction band edge as [161]

$$n(E) = N_c \exp[-(E - E_c)/kT] \tag{4.87}$$

or [80]

$$n(E) = n_s \exp[-(E - E_c)/kT] \tag{4.88}$$

where N_c is the effective density of conduction band states, Eq. E.15 and n_s is the total electron concentration. Both of these results are easily seen to be in error by integrating over all energies above the conduction band, which does not yield n_s. The total electron concentration is given by

$$n_s = \int_{E_c}^{\infty} g(E)f(E)dE \tag{4.89}$$

Hence, the correct expression for $n(E)$ in the Boltzmann approximation is

$$\begin{aligned} n(E) &= g(E)f(E) && (4.90)\\ &= g(E)e^{-(E-E_F)/kT} \\ &= \left\{\frac{\sqrt{2}m^{3/2}}{\hbar^3}(E - E_c)^{1/2}e^{-(E_c-E_F)/kT}\right\} e^{-(E-E_c)/kT} && (4.91) \end{aligned}$$

While Eq. 4.87 captures the dominant energy dependence of the exponential in Eq. 4.91, the quantity in brackets is not equal to N_c or n_s and has a different dependence on energy and temperature. The Fermi level E_F in Eq. 4.91 also depends on temperature, a fact which is entirely neglected in Eq. 4.87. Furthermore, this entire approach, including the correct expression Eq. 4.91, depends on a Boltzmann approximation to the Fermi function which is not valid under most bias conditions of interest.

The formulation presented in this and the previous chapter properly accounts for the energy and temperature dependence of the carrier distribution. If the tunneling is elastic, it is the density of states at a given energy $g_c(E)$, rather than the integrated "effective" density N_c which is of interest. Unlike N_c, $g_c(E)$ is very weakly temperature dependent [Eqs. 3.24, E.15]. Thus, the temperature dependence in Eq. 4.90 enters primarily through the Fermi function $f(E)$.

The amplitude of the threshold voltage shift in Eq. 4.37 is driven by the change in Fermi functions between stress and relaxation,

$$\Delta f(E) = f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x) \tag{4.92}$$

The Fermi functions, Eqs. 4.18 and 4.19, depend on temperature in two ways – the width of the step at the Fermi level E_F broadens with increasing temperature and the energy of the Fermi level decreases.

Figure 4.16(a) plots the functions f and f_0 for a trap at the interface at the conduction band edge ($E = 0$). For this example the initial gate bias was $V_G = 2V$ and a 5V pulse was applied during stress. The difference $\Delta f = f - f_0$ is shown and resembles a broadened impulse of width proportional to $E_F - E_{F0}$. Fig. 4.16(b) shows the function $\Delta f(E)$ for the above example at $-55°$K and 125°K. With increasing temperature, the function Δf moves lower in energy and its width broadens. Oxide traps with energies that fall in regions where Δf is nonzero undergo a change in occupancy during stress while traps at energies where $\Delta f = 0$ are either always empty or always full and therefore do not contribute to the threshold voltage shift. Thus, the function Δf acts as a "probe" which selects those trap energies which contribute to the threshold voltage shift. The contribution of a particular trap to the threshold voltage shift may either increase or decrease with temperature depending on the energy of the trap. In the example of Fig. 4.16(a), Δf is higher at -55 than at 125 and ΔV_T will therefore have a negative temperature coefficient for a trap located just above the conduction band. The solid line in Fig. 4.16(b) plots $\Delta f(125°C) - \Delta f(-55°C)$ which indicates the polarity of the temperature coefficient as a function of energy for a trap at the interface. In this example, the contribution to the temperature coefficient of ΔV_T by traps at the interface with $-0.2\text{eV} < E_t < -0.04\text{eV}$ is positive while the contribution by traps with $-0.04\text{eV} < E_t < 0.12\text{eV}$ is negative. It will be shown in Chap. 7 that NMOS transistors generally have trap densities which increase rapidly at energies above the silicon conduction band. For this reason, the threshold voltage

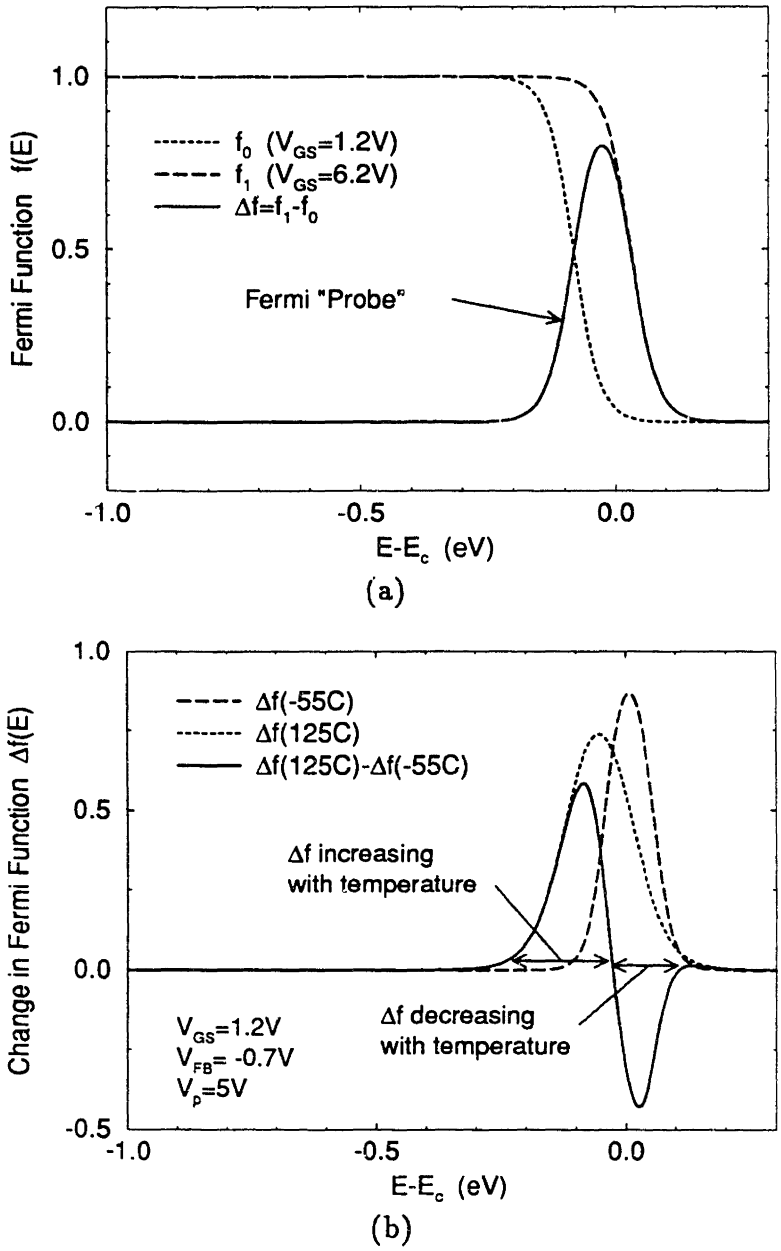


Figure 4.16: (a) The functions f and f_0 and the difference $\Delta f = f - f_0$ for a trap at $x = 0$ (b) The function $\Delta f(E) = f(E) - f_0(E)$ as a function of temperature.

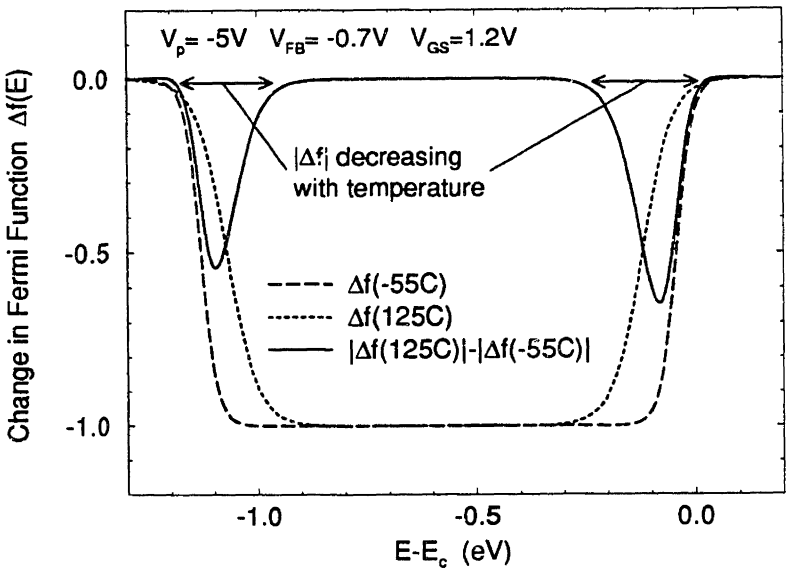


Figure 4.17: The function $\Delta f(E) = f(E) - f_0(E)$ for negative stress voltages.

shifts generally decrease with increasing temperature.

The same argument applies to negative gate voltage pulses. Fig. 4.17 shows that traps near the band edges see a change in Fermi function Δf during stress which decreases in magnitude as the temperature rises. The contribution due to traps in these regions will therefore result in a threshold voltage shift which decreases in magnitude with increasing temperature.

Due to band bending in the oxide, the range of trap energies which contribute to positive and negative temperature coefficients of the threshold voltage shift will vary with depth into the oxide and the temperature coefficient of a trap at a given energy may change depending on its depth. Thus, the temperature dependence of the trapping-induced threshold voltage shift depends strongly on the trap distribution in both energy and depth, as well as on bias conditions. The various contributions to the threshold voltage shift; *it/it*, *it/cb*, and *cb/cb* may have quite different dependences and the net temperature coefficient may be either positive or negative depending on trap distributions and bias conditions. The full temperature dependence is not obvious and must be obtained by simulation.

4.5.3 Two-Step Process

The time constant for the two step process can be approximated, to first order, as the sum of a tunneling term and a thermal term as given in Eqs. 3.102 and 3.100. In strong inversion, the thermal term is negligible so that the process is tunneling-limited with a temperature-independent time constant. Under negative gate voltage

stress, however, the discharging of oxide traps can become thermally-limited. For example, in depletion, the SRH time constant can be quite long, $\tau_{it} \gg \tau_{ot/it}$, and strongly temperature dependent [Eq. 3.71]. Electrons may remain trapped in fast surface states at low temperatures, preventing electrons in oxide traps from tunneling out.

The traps which contribute to the observed threshold voltage shifts are those which have time constants comparable to time scales of the stress and measurement. For example, at stress times on the order of 10ms, as used in our experiments, and observation intervals up to 100ms, it is only traps with time constants in a range of about 1-100ms that are important. Assuming $\tau_o = 10^{-10}$s and $x_o = 1\mathring{A}$, these traps are limited to depths of about 16-20$\mathring{A}$ from the interface. At high temperatures, the time constants for traps closer to the surface than this are sufficiently short that they are not observable in our measurements [Sec. 2.1.3].

For negative pulses which drive the surface into weak inversion or depletion, the thermal time constants for traps near the interface can become large relative to the tunneling time constant. Because τ_{it} is inversely proportional to the hole concentration p_s [Eq. 3.100], these traps may not have time to discharge during a negative gate voltage pulse until the pulse is sufficiently large to accumulate the surface. At this point, fast surface states can rapidly empty, enabling oxide traps to discharge through them, resulting in a negative threshold voltage shift. This effect becomes more pronounced at lower temperatures, since the thermal time constants grow exponentially with temperature when $n_s \ll n_1$ and $p_s \ll p_1$ [Sec. 2.1.3]. In our example, traps at depths less than 16$\mathring{A}$ may then have time constants sufficiently large to fall within the time-constant window of our measurements. With decreasing temperature, more traps have time constants which fall within the observable range and the threshold voltage shift will increase. Experimental data in support of this effect will be shown in Chap. 7.

4.5.4 Lattice Relaxation

The time constant for lattice relaxation is similar to the thermal term in the two-step process, where the term n_1 in the denominator becomes exponentially dependent on temperature in depletion. In addition, the lattice relaxation time constant [Eq. 3.99], depends on the total surface concentration of electrons and contains a thermally-activated term related to the barrier height ΔE_B. For negative gate voltage pulses, we therefore expect the lattice relaxation mechanism to have a temperature dependence similar to that of the two-step process. For positive pulses, in the extrinsic temperature regime, n_s is not strongly dependent on temperature, so the primary effect arises from the $e^{\Delta E_B/kT}$ term. This leads to a strong increase in ΔV_T with temperature, as shown by simulations in Sec. 7.2.4.

4.5.5 Summary

In summary, the net temperature dependence of the equivalent gate-source voltage shift depends on the polarity of the applied stress voltage and the mechanisms which contribute to charge trapping. In strong inversion, with positive gate voltage pulses, the movement of the Fermi probe towards lower energies will cause ΔV_{GS} to decrease if the trap density increases with energy above the conduction band as is normally the case. Due to the dependence on trap distribution, however, it is also possible for the opposite temperature coefficient to occur. Correlated mobility fluctuations cause ΔV_{GS} to decrease with increasing temperature, provided that the mobility term in Eq. 4.15 is large compared to the threshold voltage term. The thermally-activated lattice relaxation time constant decreases with increasing temperature, tending to increase ΔV_{GS} as demonstrated in Chapter 7.

For negative gate voltage pulses, the temperature dependence arises through correlated mobility fluctuations as well as through the activated terms in the lattice relaxation and two-step time constants. Decreasing the temperature increases the time constants of traps near the interface, bringing more traps into the window observable by our measurements. These traps cannot respond to fast gate voltage pulses until accumulation is reached and the large hole concentration decreases their time constants sufficiently. This leads to a negative gate voltage shift which increases with decreasing temperature.

4.6 Geometry Dependence

Gate Oxide Thickness

As the oxide thickness is scaled down together with all voltage levels such that the electric fields remain constant, it is clear from Eqs. 4.68, 4.77 and 4.81, that the threshold voltage shifts also scale down proportional to t_{ox}. Scaling will therefore have no effect on the relative error $\Delta V_T/\Delta V_G$ of the trapping-induced threshold voltage shift. Obviously, if the oxide thickness is scaled down at constant voltage levels, the threshold voltage shift will increase due to the larger electric fields which increase the range of trap energies ΔE influenced during a given gate voltage pulse.

Channel Length and Width

Since the trapped charge is very close to the interface, the factor $(t_{ox} - x)/\epsilon_{ox}$ in Eq. 4.37 is approximately equal to $1/C_{ox}$. Thus, assuming the oxide trap density $D_{ot}(E, x)$ is a function only of energy E and depth x into the oxide, the threshold voltage shift during stress is proportional to the change in trapped charge per unit area divided by the oxide capacitance per unit area. This result implies that the trapping-induced threshold voltage shifts should be independent of device area.

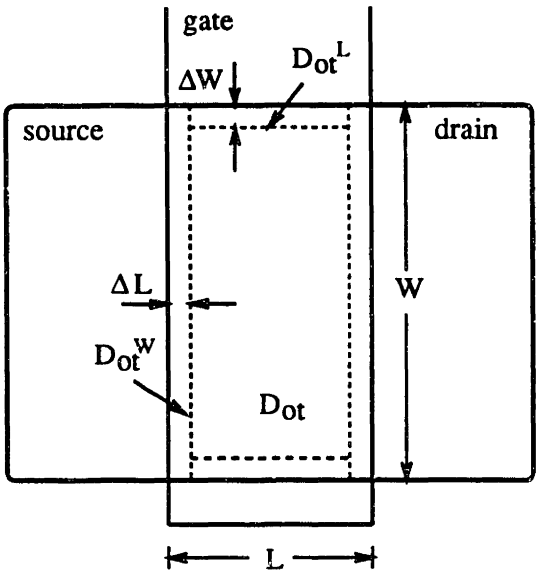


Figure 4.18: Plan view of a MOSFET defining terms used in derivation of geometry dependence

The assumption that the trap density D_{ot} is uniform in the plane may fail under certain conditions. For example, the oxide trap distribution may vary along the length of the channel as a result of the generation of oxide traps by hot carriers created by impact ionization at the drain end of the channel in saturation. The MOSFETs used in this study are virgin devices which have never been subject to hot carrier stressing of any kind, and are only operated in the linear region during measurement. Since there are no other asymmetries in device structure or biasing, there is no reason to believe the trap distribution to have any systematic dependence along the width or length dimensions and we therefore assume a uniform trap distribution over the active device area, depending only on x.

There remains, however, the possibility of edge effects due to variations in the trap density along the perimeter of the gate due to differences during processing. An enhancement of the trap distribution in these areas may result from damage created during implantation of the source-drain regions or may be due to higher trap densities in the field oxide which forms the bird's beak. In order to investigate the consequences of a variation in trap density around the periphery of the gate oxide, consider the simple model shown in Fig. 4.18. Assume a uniform trap density D_{ot} ($\mathrm{cm}^{-3}\ \mathrm{eV}^{-1}$) in the interior region of the gate oxide, far removed from the edges. Consider a strip of width ΔW along the bird's beak in which the oxide trap density is ${D_{ot}}^{L}$ ($\mathrm{cm}^{-3}\ \mathrm{eV}^{-1}$) and a strip of thickness ΔL bordering on the source-drain regions in which the trap density is ${D_{ot}}^{W}$ ($\mathrm{cm}^{-3}\ \mathrm{eV}^{-1}$). Writing Eq. 4.37 in terms of the total trapped charge

divided by the total oxide capacitance,

$$\Delta V_{Ti/j}(t_s,t_r) = \frac{q}{C_{ox}LW}\int_0^{t_{ox}}\int_{E_1}^{E_2}\left(1-\frac{x}{t_{ox}}\right)\Delta f[1-R_i(t_s)]R_j^o(t_r) \\ [D_{ot}(W-2\Delta W)(L-2\Delta L)+2W\Delta L{D_{ot}}^W+2\Delta W(L-2\Delta L){D_{ot}}^L]dE_t dx$$

For $\Delta W \ll W$ and $\Delta L \ll L$, this can be written

$$\Delta V_{Ti/j}(t_s,t_r) = \frac{q}{C_{ox}}\int_0^{t_{ox}}\int_{E_1}^{E_2}\left(1-\frac{x}{t_{ox}}\right)\Delta f[1-R_i(t_s)]R_j^o(t_r) \\ D_{ot}(E_t,x)\left[1+\gamma_w\frac{2}{L}+\gamma_l\frac{2}{W}\right]dE_t dx \qquad (4.93)$$

where γ_w and γ_l are the enhancement factors for the oxide trap densities bordering on the source-drain and bird's beak regions, respectively.

$$\gamma_w = \Delta L\frac{{D_{ot}}^W}{D_{ot}} \qquad (4.94)$$

$$\gamma_l = \Delta W\frac{{D_{ot}}^L}{D_{ot}} \qquad (4.95)$$

Assuming ΔL and ΔW fixed by processing conditions, γ_w and γ_l are just constants and the geometrical factors can be removed from the integral in Eq. 4.93. Thus, Eq. 4.93 predicts that an enhanced oxide trap density along the periphery of the gate will cause the threshold voltage shift to vary inversely with the gate length and width,

$$\frac{\Delta V_T(L,W)}{\Delta V_{T0}} = 1+\frac{2\gamma_w}{L}+\frac{2\gamma_l}{W} \qquad (4.96)$$

where ΔV_{T0} is the threshold voltage shift for L and W sufficiently large that edge effects are unimportant, which is the result predicted by the standard theory.

4.7 Hysteresis Effects

When a MOSFET is stressed with a large gate voltage for a period of time t_s, it is observed that the resulting threshold voltage shift relaxes with a very long tail which may persist for times t_r which greatly exceed t_s. Because the effect persists for a time much longer than the time required to initiate it, we have previously referred to this effect as "threshold voltage hysteresis" [10, 62].

Consider Eq. 3.96 in the case of an initially empty oxide trap, so that at $t = 0^-$, $f_{ot}(E,x,0^-) = 0$. Assume that, following the application of a gate voltage step, the Fermi level at the interface rises above the trap energy so that the oxide trap begins to charge. While the exact mechanism by which the trap charges is not of interest

in the present discussion, the oxide trap time constant is assumed much larger than the equilibration times for either interface states or inversion layer charge carriers so that, immediately following stress, the initial conditions can be taken to be

$$\begin{aligned} f_{ot}(E,x,0^+) &= f_{ot}(E,x,0^-) = 0 \\ f_{ot}(E,x,\infty) &= 1 \end{aligned}$$

Thus, the initial rate at which oxide traps fill is

$$\left.\frac{dn_{ot}}{dt}\right|_{t=0^+} = \frac{N_{ot}}{\tau_f} \tag{4.97}$$

where τ_f represents the value of either $\tau_{ot/cb}$ or $\tau_{ot/it}$ during stress.

Following a stress period of duration t_s, oxide traps fill to a fractional occupancy $f_{ot}(E,x,t_s)$. When the gate voltage returns to its initial value, the Fermi level at the interface falls and the initial conditions for the discharging of traps are

$$\begin{aligned} f_{ot}(E,x,0^-) &= f_{ot}(E,x,t_s) \\ f_{ot}(E,x,\infty) &= 0 \end{aligned}$$

Thus, the initial rate at which traps empty, after a stress time t_s is

$$\left.\frac{dn_{ot}}{dt}\right|_{t=0} = -\frac{n_{ot}(t_s)}{\tau_e} \tag{4.98}$$

where τ_e represents the value of the time constant at the bias condition following stress.

For short stress times, $n_{ot}(t_s) \ll N_{ot}$, the initial rate at which traps empty is much slower than the rate at which they are filled, even for symmetrical time constants $\tau_f \sim \tau_e$. If an NMOS transistor is stressed for a given period of time, t_s and then allowed to relax for an equal time, some residual charge Δn_{ot} will remain trapped in the oxide, which may manifest itself as hysteresis in circuit applications. While the long relaxation time relative to the stress time may be enhanced by inherent asymmetries between between the time constants during capture and emission, $\tau_f < \tau_e$, the observed hysteresis may simply be due to the fact that the charging transient for many of the oxide traps is interupted before going to completion, $n_{ot}(t_s) < N_{ot}$. An asymmetry in rates, but not necessarily in time constants, is the condition required for hysteresis.

The observed hysteresis may be due to a combination of these effects, each of which is discussed below. In the following we shall use the term relaxation rather than hysteresis in order to deemphasize any implied asymmetry in the underlying physical mechanism, which is not a necessary condition for the observation of the hysteresis effect.

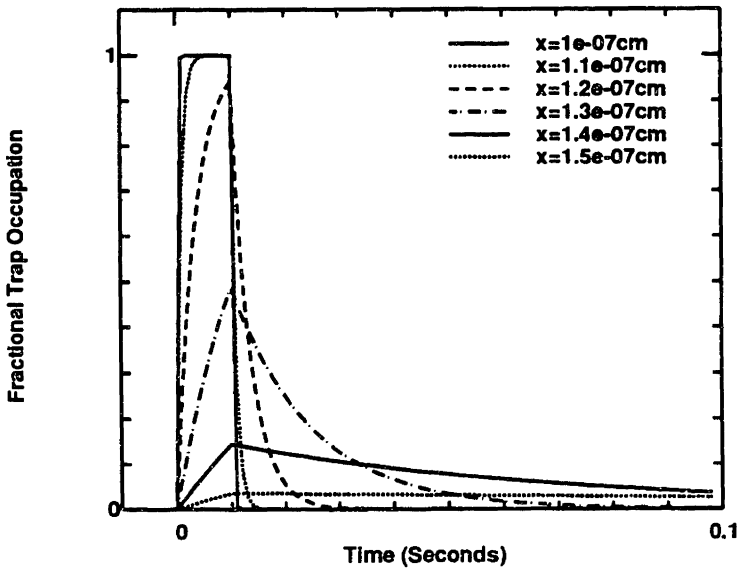


Figure 4.19: Fractional occupation of oxide traps $g(x,t_s)[1-g(x,t_r)]$ at various depths as a function of time during stress and emission

4.7.1 Charging and Emission Kinetics

Figure 4.19 shows the fractional occupation of a single oxide trap, $g(x,t_s)[1-g(x,t_r)]$, at various depths from the interface as a function of time during and after a stress period of 10ms using Eqs. 4.58 and 4.59. For traps less than about $10\mathring{A}$ from the interface, the filling transient goes to completion during the stress time and the emission transient is symmetrical. Deeper traps, with correspondingly longer time constants, are only partially filled during the stress period and hence the rate at which they discharge is much less than the capture rate, leading to a long relaxation time. The total response is the superposition of the response of traps at all depths,

$$\Delta V_T(t_s,t_r) = \frac{qD_{ot}}{C_{ox}} \int_0^{t_{ox}} \left(1 - \frac{x}{t_{ox}}\right) g(x,t_s)[1-g(x,t_r)]dx \qquad (4.99)$$

Equation 4.99 is integrated numerically and plotted in Fig. 4.20. The dotted line shows the path the charging transient would take were it not interupted after $t_s =$ 10ms. The figure emphasizes the fact that the initial amplitude of the threshold voltage shift may be very much larger than the long tail which is actually observed. The initial discharge of fast surface states and near-interface oxide traps is very rapid as shown in Figs. 4.20 and 4.9. This initial transient is too fast to be measured by the test circuit described in Sec. 6.2.2, so that what is actually observed is the long-tail portion of the emission transient in Fig. 4.20, with peak amplitude on the order of one millivolt, which persists for periods up to hundreds of milliseconds.

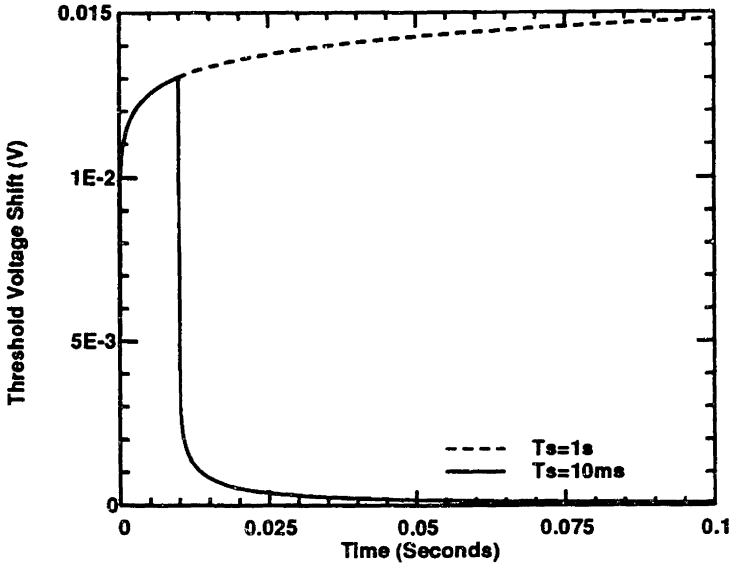


Figure 4.20: Threshold voltage shift as a function of stress time and relaxation time for $t_{ox} = 250\mathring{A}$ and $D_{ot} = 1 \times 10^{17}$ cm^{-3}.

4.7.2 Time Constant Asymmetries

Differences in Capture and Emission Mechanisms

The exact model presented in this thesis is inherently asymmetrical, primarily due to the fact that, for a given oxide trap, the capture and emission processes may occur by different mechanisms. For example, for traps which fall into the it/cb category, if $\tau_{ot/it} < \tau_{ot/cb}$, the capture rate will exceed the emission rate in Eq. 4.42. Even when the capture and emission rates occur by the same mechanism (e.g. cb/cb and it/cb), the time constants in Eqs. 3.18 and 3.43 depend on energy and electric field and hence differ during stress and relaxation. These effects are built into the exact model and implemented in the simulation program.

Differences in Barrier Height

Small changes in the height of the tunneling barrier occur due to changes in the oxide electric field during stress and emission as mentioned above. Larger changes in barrier height may occur due to defect relaxation. If, after capture of an electron, the energy of the trapped electron decreases, then the trapped electron will see a higher energy barrier during emission than during capture. Relaxation energies of as much as 1 eV have been observed [30]. Figure 4.21(b) shows the effect on Eq. 4.99 of an increase in

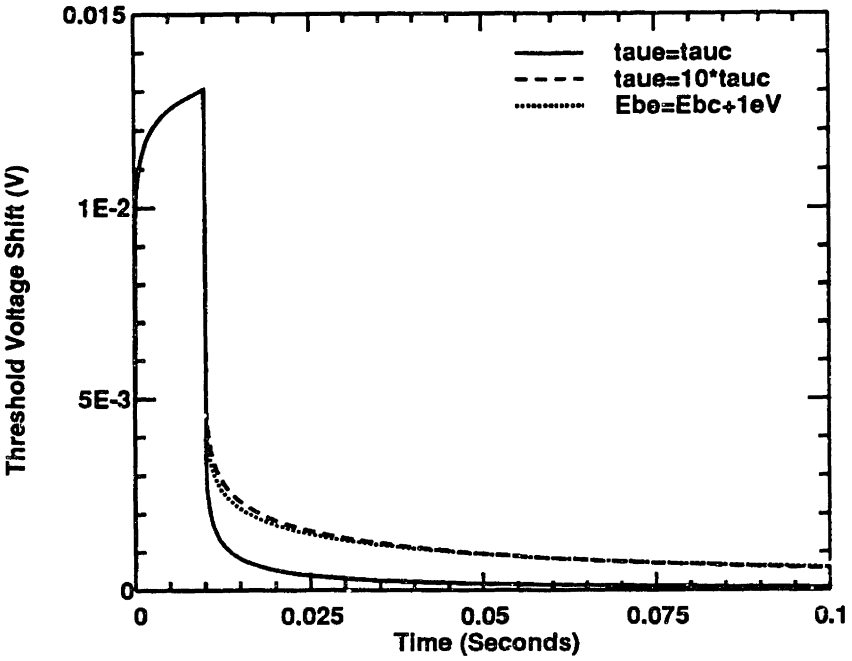


Figure 4.21: Charging and Emission transients obtained by integrating Eq. (a) Solid line: $\tau_f = \tau_e$ (b) Dashed line: $\tau_e = 10\tau_f$ (c) Dotted line: $E_{Be} = E_{Bc} + 1eV$

the barrier height during emission by 1 eV, $E_{Be} = E_{Bc} + 1\text{eV}$, with

$$\tau = \tau_o \exp\left(2\sqrt{\frac{2mE_B}{\hbar^2}}x\right) \tag{4.100}$$

It can be seen that a 1eV difference in barrier height between capture and emission gives rise to about a factor of 10 difference in time constants, which could lead to large hysteresis effects.

Inelastic Effects

In elastic tunneling from the conduction band (Sec. 3.4.1), the tunneling rate does not depend on the total inversion layer charge concentration, but only on the carrier concentration at the same energy as the trap. In this model, the carrier concentrations do not enter into the time constant. In the trap-assisted tunneling process, the carrier concentrations enter indirectly, through the SRH time constant τ_{it}. However, at room temperature and in strong inversion, this time constant is normally so much less than the tunneling time constant, that changes in surface carrier concentration will have a negligible effect.

In Sec. 3.4.3 it was shown that if oxide trap energies relax by multiphonon emission following an elastic tunneling event, then the time constant depends inversely on the

surface carrier concentration [Eq. 3.29]. Consider the lr/lr component. In strong inversion, the time constant during capture is given by Eqs. 4.51 and 4.84,

$$\begin{aligned}\tau_f(x) &= \tau_{oc}e^{x/x_o} \\ &= \frac{e^{\Delta E_B/kT}}{\alpha(V_{GS}-V_T)}e^{x/x_o}\end{aligned}$$

During emission, the MOSFET is still in strong inversion and the time constant is given by Eqs. 4.52 and 4.85,

$$\begin{aligned}\tau_e(x) &= \tau_{oe}e^{x/x_o} \\ &= \frac{e^{\Delta E_B/kT}}{\alpha(V_{GS}^o-V_T)}e^{x/x_o}\end{aligned}$$

Assuming $V_T \approx 1V$, $V_{GS}^o = 2V$ and an applied pulse of amplitude $V_p = V_{GS} - V_{GS}^o = 4V$, the time constant during emission is 5 times longer than during capture. From Eqs. 4.60 and 4.61, this means that after a relaxation time equal to the stress time, all the filled traps have not been emptied, but remain filled over a depth

$$x_f - x_e = x_o \ln\left(\frac{t_s}{\tau_o}\right) - x_o \ln\left(\frac{t_s}{5\tau_o}\right) = 1.609x_o \approx 1.609\mathring{A} \qquad (4.101)$$

The effect of a time constant that is an order of magnitude larger during stress than during emission is shown in Fig. 4.21 and results in a substantial increase in the magnitude and duration of the transient threshold voltage shift.

4.8 Dynamic Effects

It was shown above that the threshold voltage relaxation may manifest itself in the form of hysteresis under conditions in which a MOSFET is stressed for a short time and then allowed to relax for a longer time. In clocked circuits, a MOSFET may be subjected to periodic gate voltage pulses. Under such dynamic stress conditions the threshold voltage shift may exhibit a slow drift due to charging and discharging transients which cannot go to completion during a clock period. Figure 4.22 shows the simulated threshold voltage shift in a MOSFET stressed with 5V pulses at 1MHz with a 50% duty cycle. The curves are obtained from a generalization of Eq. 4.99,

$$\Delta V_T(t) = \begin{cases} \frac{qD_{ot}}{\epsilon_{ox}} \int_0^{t_{ox}} (t_{ox}-x)(1-e^{-/\tau})^{\frac{n}{2}}(e^{-T/\tau})^{\frac{n}{2}}(1-e^{-(t-nT)/\tau})dx & n \text{ even} \\ \frac{qD_{ot}}{\epsilon_{ox}} \int_0^{t_{ox}} (t_{ox}-x)(1-e^{-T/\tau})^{\frac{n+1}{2}}(e^{-T/\tau})^{\frac{n-1}{2}}e^{-(t-nT)/\tau}dx & n \text{ odd} \end{cases} \qquad (4.102)$$

where $2T$ is the period and $n = 0, 1, 2, \ldots$.

This slow threshold voltage drift will eventually reach a steady state, however it may be a source of error in certain types of analog circuits under transient or start-up conditions, or when the amplitude of the stress voltage changes.

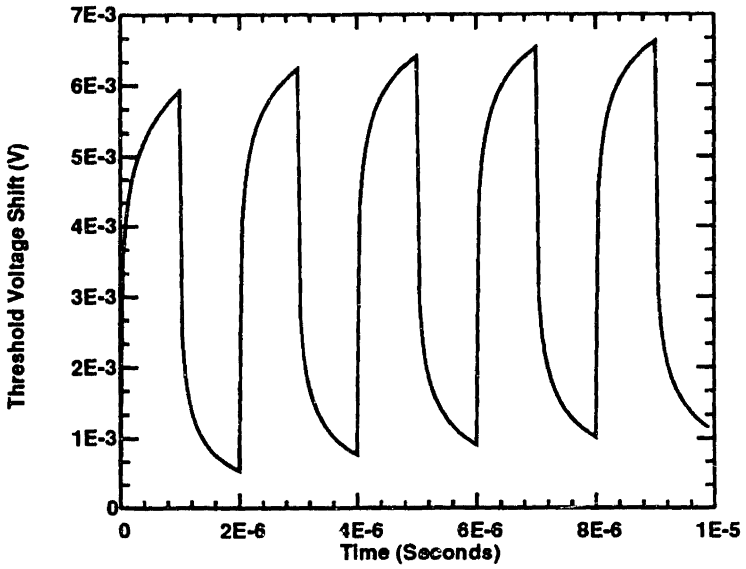


Figure 4.22: Threshold voltage shift as a function of time for 1MHz pulse rate

4.9 Summary

The general equation for the equivalent gate-source voltage shift due to tunnel exchange with oxide traps under large-signal stress can be written

$$\Delta V_{GS} = \sum_{i,j} \Delta V_{GS(i/j)} \tag{4.103}$$

where $i,j \in \{cb, vb, it, lr\}$ represent tunneling from conduction band ($j = cb$) and valence band ($j = vb$) states, interface-state assisted tunneling ($j = it$) and tunneling/lattice relaxation ($j = lr$). The component due to oxide traps which charge by process i during stress and discharge by process j during emission is

$$\begin{aligned}\Delta V_{GS(i/j)}(t_s, t_r) \;=\; & \frac{q}{C_{ox}} \int_0^{t_{ox}} \int_{E_1(i,j)}^{E_2(i,j)} \left[1 - \frac{x}{t_{ox}} + \mu_N S_o\right] [f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x)] \\ & \times [1 - R_i(E_t + q\mathcal{E}x, x, t_s)]\, R_j^o(E_t + q\mathcal{E}_o x, x, t_r) D_{ot}(E_t, x)\, dE_t\, dx\end{aligned} \tag{4.104}$$

where $D_{ot}(E_t, x)$ is the density of traps with ground state energy E_t at depth x from the interface. $(1 - R_i)$ and R_j^o represent the charging and emission transients during stress and relaxation respectively. In general, these may be have either first-order (band-to-trap tunneling) or second-order (two-step process) dynamics. In a simple first-order approximation [Sec. 3.7], as used in the simulations of Sec. 7.2.4, these relaxation transients become

$$R_i(E, x, t) = e^{-t/\tau_{ot/i}(E,x)}$$

$$R_j^o(E,x,t) = e^{-t/\tau_{ot/j}^o(E,x)}$$

The time constants during stress, $\tau_{ot/i}$, and during relaxation, $\tau_{ot/j}^o$, are evaluated at the bias conditions during and after stress, respectively. The first-order time constants time constants $\tau_{ot/j}$ are summarized in Table 4.5 for each of the processes that we have considered. The transformation $E \rightarrow E_t + q\mathcal{E}x = E_t - \frac{qV_{ox}x}{t_{ox}}$ has been used to incorporate oxide band bending. All trap energies are measured upward with respect to an origin ϕ_c below the oxide conduction band, as shown in Fig 4.1. The limits of integration in Eq. 4.104 depend on energy and depth and are summarized in Tables 4.1–4.4.

Table 4.6 shows the forms of the time constants in the tunneling-limited (either n_s or p_s large relative to n_1 and p_1) and thermally-limited (n_s and p_s negligibly small relative to n_1 and/or p_1) regimes. In an NMOS transistor, the former case applies to positive (inverting) and large negative (accumulating) stresses, while the latter occurs for negative pulses which drive the surface into weak inversion or depletion. Temperature dependence enters the time constants through the term $\exp(\Delta E_B/kT)$ for lattice relaxation and through the activated factors n_1 and p_1. In the bias regions where these terms are important – weak inversion and depletion – the carrier concentrations are small, so that the Boltzmann approximations of Eq. 3.53 and 3.54 are valid. In terms of the trap energies E_t, referenced to the conduction band edge E_c,

$$\begin{aligned} n_1 &= N_c e^{(E_t+q\mathcal{E}x)/kT} \\ p_1 &= N_v e^{-(E_t+q\mathcal{E}x+E_G)/kT} \end{aligned}$$

The strength of correlated mobility fluctuations is determined by the scattering rate $S(x)$. If mobility fluctuations are negligible relative to number (threshold voltage) fluctuations, then Eq. 4.104 is just the threshold voltage shift. In the model of Eq. 4.14, $S(x) = S_o/C_{ox}(V_{GS} - V_T)$ and the dependence on gate voltage vanishes. The scattering rate S_o has been assumed constant, but may also depend on the trap energy, for example if the charge character (donor-like or acceptor-like) of the trap depends on energy [157]. Correlated mobility fluctuations can be incorporated into an effective threshold voltage shift, of the form

$$\Delta V_T' = \frac{qt_{ox}'}{\epsilon_{ox}} \int_0^{t_{ox}'} \int_E \left(1 - \frac{x}{t_{ox}'}\right) \Delta n_{ot}(E,x,t)\, dE\, dx \tag{4.105}$$

where the trapped charge Δn_{ot} is calculated in the usual manner, with an oxide thickness t_{ox}, but the transformation from trapped charge to effective threshold voltage shift uses an effective oxide thickness given by

$$t_{ox}' = (1 + \mu_N S_o) t_{ox} \tag{4.106}$$

For $x \ll t_{ox}$, correlated mobility fluctuations therefore enter the equivalent gate voltage shift as a constant (temperature-dependent) scale factor on the threshold voltage shift.

The attenuation constants in the oxide, as given in Appendix B, written in terms of the energy E_t of a trap at x are

$$K_c(x') = \left\{\left(\frac{2M_c^*}{\hbar^2}\right)[\phi_c - E_t - q\mathcal{E}(x-x')]\right\}^{1/2}$$
$$K_v(x') = \left\{\left(\frac{2M_v^*}{\hbar^2}\right)[\phi_v + E_G + E_t + q\mathcal{E}(x-x')]\right\}^{1/2}$$

The attenuation constant κ in the semiconductor bandgap is, in a two-band model [Eq. 3.95],

$$\kappa = \frac{\sqrt{2}}{\hbar}\left[\frac{m_c^* m_v^*(E_t + q\mathcal{E}x)(E_t + E_G + q\mathcal{E}x)}{m_c^*(E_t + q\mathcal{E}x) - m_v^*(E_t + E_G + q\mathcal{E}x)}\right]^{1/2}$$

The wavevectors in the semiconductor are

$$k_c = \left[\frac{2m_c^*}{\hbar^2}(E_t + q\mathcal{E}x)\right]^{1/2}$$
$$k_v = \left[\frac{-2m_v^*}{\hbar^2}(E_t + E_G + q\mathcal{E}x)\right]^{1/2}$$

Note that, since the last expression is only evaluated in the valence band ($E_t + q\mathcal{E}x < E_G$), the quantity inside the square root is positive.

Approximate expressions were derived for the three dominant threshold voltage components in an NMOS transistor in strong inversion, assuming mobility fluctuations to be negligible. These are summarized below for the two-step process:

Interface State/Interface State (it/it):

$$\Delta V_{Tit/it}(t_s, t_r) = \frac{qD_{ot}x_o}{C_{ox}}(E_F - E_{F0})\ln\left(1 + \frac{\tau_{it0}t_s}{\tau_{it1}t_r}\right)$$

Interface State/Conduction Band (it/cb):

$$\Delta V_{Tit/cb} = \frac{q^2 D_{ot}x_o^2}{\epsilon_{ox}}\Delta V_{ox}\ln\left(1 + \frac{\tau_{cb0}t_s}{\tau_{it1}t_r}\right)\left[\gamma + \frac{1}{2}\ln\left(1 + \frac{\tau_{cb0}t_s}{\tau_{it1}t_r}\right) + \ln\left(\frac{t_r}{\tau_{cb0}}\right)\right]$$

Conduction Band/Conduction Band (cb/cb):

$$\Delta V_{Tcb/cb}(t_s, t_r) \approx \frac{qD_{ot}x_o}{C_{ox}}kT\ln\left\{1 + \exp\left[\frac{E_F - E_c}{kT}\right]\right\}\ln\left(1 + \frac{\tau_{cb0}t_s}{\tau_{cb1}t_r}\right)$$

In this first-order approximation, the only difference between the two-step and lattice

Table 4.5: Summary of First-Order Time Constants

$\tau_{ot/j}(E_t, x)$ **(TIME CONSTANT)**	$\tau_o(E_t, x)$
INTERFACE STATE-ASSISTED ($j = it$)	EXACT:
$\frac{1}{\sigma_{no}\overline{v}(n_s+n_1)+\sigma_{po}\overline{v}(p_s+p_1)} + \tau_o(E_t, x)e^{\lambda(E_t,x)}$	
[Eqs. 3.43, 3.71, 3.83]	APPROXIMATION:
LATTICE RELAXATION ($j = lr$)	EXACT:
$\tau_o(E_t, x)e^{\lambda_c(E_t,x)}$ [Eq. 3.27]	APPROXIMATION:
CONDUCTION BAND TUNNELING ($j = cb$)	EXACT:
$\tau_o(E_t, x)e^{\lambda(E_t,x)}$	
[Eq. 3.18]	APPROXIMATION:
VALENCE BAND TUNNELING ($j = vb$)	EXACT:
$\tau_o(E_t, x)e^{\lambda(E_t,x)}$	
[analogous to Eq. 3.18]	APPROXIMATION:

(TIME CONSTANT PREFACTOR)	$\lambda(E_t, x)$ (WKB EXPONENT)
$\frac{M^{*2}}{4\pi\hbar^3\sigma_n D_{it}(E_t+q\mathcal{E}x)}\left[\frac{\kappa+K(0)}{\kappa K^2(0)K^2(x)}\right]$ (1-d) [Eq. B.114] $\frac{M^{*2}}{8\pi^2\hbar^3 D_{it}(E_t+q\mathcal{E}x)}\left[\frac{\kappa+K(0)}{\kappa K^2(0)}\right]$ (3-d) [Eq. B.125]	$\frac{4}{3q\mathcal{E}}\left(\frac{2M_c^*}{\hbar^2}\right)^{1/2}$ $\times[\xi(\phi_c-E_t)-\xi(\phi_c-E_t-q\mathcal{E}x)]$ $\xi(u)=\begin{cases} u^{3/2} & \text{(one-band)} \\ u^{3/2}-\frac{3}{10}\frac{M_c^*}{M_v^*}\frac{u^{3/2}}{E_{Gb}} & \text{(two-band)}\end{cases}$ [Eq. 3.92]
constant $\begin{cases}\tau_{it0} & \text{(during relaxation)} \\ \tau_{it1} & \text{(during stress)}\end{cases}$	$2Kx \equiv x/x_o$
$[\sigma_n e^{-\Delta E_{Bc}/kT}\overline{v}(n_s+n_1)$ $+\sigma_p e^{-\Delta E_{Bv}/kT}e^{\lambda_v/\lambda_c}\overline{v}(p_s+p_1)]^{-1}$ [Eq. 3.34]	$\lambda_c=\frac{4}{3q\mathcal{E}}\left(\frac{2M_c^*}{\hbar^2}\right)^{1/2}[(\phi_c+q\mathcal{E}x)^{3/2}-\phi_c^{3/2}]$ $\lambda_v=\frac{4}{3q\mathcal{E}}\left(\frac{2M_v^*}{\hbar^2}\right)^{1/2}[\phi_v^{3/2}-(\phi_v-q\mathcal{E}x)^{3/2}]$ [Eqs. 3.28, 3.32]
$\frac{e^{\Delta B_B/kT}}{\alpha(V_{GS}-V_T)}$ [Eqs. 4.84, 4.86] (strong inversion)	$2Kx \equiv x/x_o$
$\frac{4\pi\hbar}{q\sigma_n\mathcal{E}K_c^2(x)}\left(\frac{M_c^*}{m_c^*}\right)^2\frac{K_c(x)-K_c(0)}{1-\exp\left\{-\frac{\hbar^2 m_c^* k_c^2[K_c(x)-K_c(0)]}{M_c^{*2}q\mathcal{E}}\right\}}$ (1-d) [Eq. B.33] $\frac{m_c^* x}{2\hbar}\frac{k_c^2+K_c^2(0)}{k_c K_c^2(0)}$ (3-d) [Eq. B.73]	$\frac{4}{3q\mathcal{E}}\left(\frac{2M_c^*}{\hbar^2}\right)^{1/2}$ $\times[(\phi_c-E_t)^{3/2}-(\phi_c-E_t-q\mathcal{E}x)^{3/2}]$ [Eq. 3.18]
constant $\begin{cases}\tau_{cb0} & \text{(during relaxation)} \\ \tau_{cb1} & \text{(during stress)}\end{cases}$	$2Kx \equiv x/x_o$
$\frac{4\pi\hbar}{q\sigma_p\mathcal{E}K_v^2(x)}\left(\frac{M_v^*}{m_v^*}\right)^2\frac{K_v(x)-K_v(0)}{1-\exp\left\{-\frac{\hbar^2 m_v^* k_v^2[K_v(x)-K_v(0)]}{M_v^{*2}q\mathcal{E}}\right\}}$ (1-d) [Eq. B.33] $\frac{m_v^* x}{2\hbar}\frac{k_v^2+K_v^2(0)}{k_v K_v^2(0)}$ (3-d) [Eq. B.73]	$\frac{4}{3q\mathcal{E}}\left(\frac{2M_v^*}{\hbar^2}\right)^{1/2}[(\phi_v+E_G+E_t+q\mathcal{E}x)^{3/2}$ $-(\phi_v+E_G+E_t)^{3/2}]$ (analogous to Eq. 3.18)
constant $\begin{cases}\tau_{vb0} & \text{(during relaxation)} \\ \tau_{vb1} & \text{(during stress)}\end{cases}$	$2Kx \equiv x/x_o$

relaxation processes is their time constants. Thus, the above equations also apply to lattice relaxation with the substitutions

$$
\begin{aligned}
it &\longrightarrow lr \\
\tau_{it1} &\longrightarrow \frac{e^{\Delta E_B/kT}}{\alpha(V_{GS}-V_T)} \\
\tau_{it0} &\longrightarrow \frac{e^{\Delta E_B/kT}}{\alpha(V^o_{GS}-V_T)}
\end{aligned}
$$

Table 4.6: Limiting Forms of Time Constants $\tau_{ot/j}(E_t, x)$

MECHANISM	TUNNELING-LIMITED ($n_s \gg p_s, n_1, p_1$) (inversion/accumulation)	THERMALLY-LIMITED ($n_s, p_s \ll n_1, p_1$) (weak inversion/depletion)
INTERFACE STATE-ASSISTED (it)	$\tau_o(E_t, x)e^{\lambda}(E_t, x)$	$\frac{1}{\overline{v}(\sigma_{no}n_1+\sigma_{po}p_1)}$
LATTICE RELAXATION (lr)	$\frac{e^{\Delta E_{Bc}/kT}\, e^{\lambda_c(E_t,x)}}{\sigma_n \overline{v} n_s}$	$\frac{e^{\lambda_c(E_t,x)}}{\overline{v}\left(\sigma_n n_1 e^{-\Delta E_{Bc}/kT} + \sigma_p p_1 e^{-\Delta E_{Bv}/kT} e^{\lambda_v/\lambda_c}\right)}$
CONDUCTION BAND TUNNELING (cb)	$\tau_o(E_t, x)e^{\lambda(E_t,x)}$	$\tau_o(E_t, x)e^{\lambda(E_t,x)}$ (temperature-independent)
VALENCE BAND TUNNELING (vb)	$\tau_o(E_t, x)e^{\lambda(E_t,x)}$	$\tau_o(E_t, x)e^{\lambda(E_t,x)}$ (temperature-independent)

Chapter 5

Dielectric Relaxation in Capacitors

5.1 Introduction

Capacitors, as well as transistors, in analog circuits are frequently subjected to large-signal voltage excursions during which charge may be exchanged with oxide traps by tunneling. If this trapping produces threshold voltage relaxation in MOSFETs then it is reasonable to expect that similar phenomena should be observable in integrated circuit capacitors. This chapter investigates the consequences of interfacial tunneling in capacitors, applying the theory developed in Chapters 3 and 4. The analysis of capacitors is greatly simplified by two distinguishing features; first, since the plates are either metal or degenerately doped material, the Fermi level lies within the conduction band and moves negligibly with voltage and temperature. Thus, only the *cb/cb* component derived in Chap. 4 contributes to the charging and discharging of oxide traps. Secondly, since there is no current flow, there is no contribution due to mobility fluctuations.

When a voltage is applied across a dielectric, several types of charge motion take place as discussed in Chap. 2. First there are the electronic and ionic polarizations which are virtually instantaneous at electrical frequencies and comprise the usual dielectric constant. A capacitor which contains only these components is essentially "ideal" and when subjected to a voltage step V through a resistance R, the normal displacement current is given by

$$I = \frac{V}{R} \exp\left(-\frac{t}{RC}\right)$$

For typical integrated circuit values as used in our measurements, e.g. $R = 10\Omega$ and $C = 100\text{pF}$, the current decays to e^{-1} of its initial value in 1ns.

In a nonideal capacitor which has a finite resistance there will also be an instantaneous conduction current of small magnitude. This may be due to ionic conduction

in some materials such as glass [162] or to electronic conduction if carriers are injected across the potential barriers at the contacts. These contributions are normally negligible in device-grade capacitors under normal circuit voltages so that the total charging current decays to an insignificant level on the order of a nanosecond.

In real capacitors there is, in addition to the large initial charging current and the small steady state conduction current, a component of intermediate magnitude which decays very slowly, persisting for times many orders of magnitude larger than RC. This is sometimes referred to as the *anomalous current*, although this term is really is misnomer since polarization is the very essence of a dielectric and the fact that some of these processes have time constants in the electrical frequency range should hardly be considered as an aberration. The better and more widely used term is *dielectric relaxation*, which refers to the time delay required for the polarizations to form after a step change in electric field, or *dielectric absorption* since this delay results in a storage of energy in the dielectric which is not immediately accessible to the outside world.

Relaxation effects can also be observed when a capacitor that has been charged to a voltage for a prolonged period of time is suddenly short-circuited. In addition to the usual impulsive discharge current, there is a current component of intermediate magnitude which decays with a very long tail, which is frequently referred to as a *depolarization current*. The same mechanisms appear under AC excitations in the form of dielectric losses and dispersion of the dielectric constant.

In circuits, the effects of dielectric relaxation usually manifest themselves in the form of voltages rather than currents. Consider a capacitor, e.g. in a sample-and-hold circuit, which is charged to a fixed voltage for some period of time, then momentarily short circuited as shown in Fig. 1.3. An ideal capacitor which is subsequently open-circuited will remain at zero volts. Due to dielectric relaxation however, a potential difference will develop across the plates of a nonideal capacitor. This voltage, which is always of the same polarity of the original charging voltage, is known as a *recovery voltage*.

Dielectric relaxation has long been recognized as a source of error in sample-and-hold circuits, analog computation and charge-redistribution analog-to-digital conversion circuits. There have been many studies of the effects of dielectric absorption on circuits, on compensation techniques and on empirical modeling [163, 164, 165, 166, 167, 168]. Dielectric relaxation has been studied in glasses [169, 170], oxides [171, 172, 173, 174] and polymers [175, 176, 177, 178, 179, 180]. There has been relatively little work, however, on dielectrics such as SiO_2 and Si_3N_4 of interest to silicon integrated circuits [181, 182]. Previous studies have focused on absorption currents, at long times and often under high voltage conditions with DC currents flowing in the insulator.

A recent study by Fattaruso et al [166] makes a significant contribution toward the lack of experimental data on insulators relevant to MOS technologies and the effects of

dielectric relaxation on circuit performance. Asymptotic recovery voltages in typical integrated circuit capacitor structures, expressed as a fraction of the stress voltage, were found to range from 20ppm in thermal SiO_2 to 1000ppm in nitrided SiO_2. The levels were also found to depend on the material used for the plates. Experiments and simulations, using Dow's model [Fig. 5.1], on charge-redistribution A/D converters showed that dielectric relaxation causes a shift in the code transition voltages which decreases with increasing delay between a sudden input transition and the conversion. For a $5\mu s$ delay, the code transition voltage error is on the order of 0.01% of full scale for a 5V input step. This study did not, however, suggest a physical mechanism for the dielectric relaxation.

The present study complements the work of Fattaruso et al, which focused on circuit modeling, by focusing on the physical mechanism responsible for dielectric relaxation at time and voltage scales relevant to analog circuit operation. We have previously suggested interfacial tunneling as a possible explanation for relaxation effects in MOS dielectrics [62]. In this chapter we apply the tunneling theory developed in the previous two chapters to the analysis of dielectric relaxation in capacitors. In particular, we derive expressions for the recovery voltage as a function of charging voltage and the duration of the stress, discharge and relaxation intervals. In Chapter 7, simulation results using the tunneling model are compared with experimental data taken from capacitor structures with conventional and nitrided SiO_2 dielectrics from the same processes used in Ref. [166].

5.2 Empirical Models

It has been known for nearly a century that the decay of absorption currents with time follows, virtually "universally" [183], the Curie-von Schweidler law,

$$I(t) = A(T)t^{-n} \tag{5.1}$$

where $A(T)$ is temperature dependent and $n \approx 1$ [179, 178].

A relaxation process characterized by a single time constant can be shown to have a complex dielectric constant of the form [184]

$$\epsilon^*(\omega) = \epsilon'(\omega) - j\epsilon''(\omega)$$

where

$$\epsilon'(\omega) = \epsilon_\infty + \frac{\epsilon_s - \epsilon_\infty}{1 + \omega^2\tau^2} \tag{5.2}$$

$$\epsilon''(\omega) = (\epsilon_s - \epsilon_\infty)\frac{\omega\tau}{1 + \omega^2\tau^2} \tag{5.3}$$

and ϵ_s and ϵ_∞ are the static and optical frequency ($10^{10} - 10^{12}$ $\sec^{-1}$) dielectric constants, respectively. Equations 5.2 and 5.3 are known as the Debye equations since they were originally derived by Debye based on a model of molecular polarization.

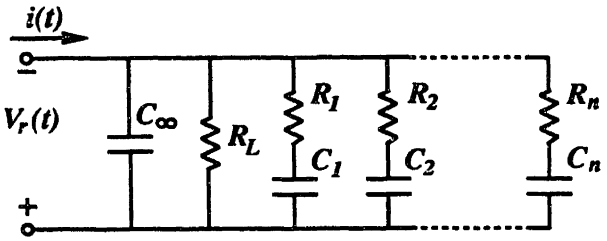


Figure 5.1: Dow's model for dielectric absorption in a capacitor.

Real dielectrics, however, differ substantially from this model. By curve fitting measured data, Cole and Cole [185, 186] found that the dielectric constant could be better approximated by the empirical expression:

$$\epsilon^*(\omega) = \epsilon_\infty + \frac{(\epsilon_o - \epsilon_\infty)}{1 + j\omega\tau_o{}^{1-\alpha}}$$

No physical mechanism was offered to explain this result.

An empirical model for a capacitor has been presented by Dow [168] which incorporates a distribution of time constants as found in real materials,

$$C^*(\omega) = C_\infty + \sum_{k=1}^{N} \frac{C_k}{1 + j\omega R_k C_k}$$

where the C_k's and R_k's are fitted to the capacitor in question over an appropriate range of frequencies. This model is useful for circuit applications because, unlike that of Cole and Cole, it can be synthesized with a finite number of lumped elements as shown in Fig. 5.1. Dow's model has been used to fit data from a variety of integrated circuit capacitor structures [166, 167].

Despite the success of these empirical modeling efforts, the physical causes of dielectric relaxation are still not completely understood. A number of mechanisms have been suggested, the most important of which are reviewed below.

5.3 Mechanisms

Dipolar Orientation

The traditional and best known explanation for dielectric relaxation is due to the orientational polarization of dipoles. This mechanism was first proposed by Debye, who applied it to the orientation, in an electric field, of polar groups in liquids [23]. In solids, dipoles may arise from point defects [187] which require some time to change state due to a particle passing over a potential barrier [177].

For a sufficiently wide distribution of relaxation times, dipole relaxations can account for the observed time dependence given by Eq. 5.1 [185, 186]. Theories based on

dipole relaxation predict an absorption current which is linearly proportional to field and thermally activated. If the dipoles are distributed homogeneously throughout the insulator, the absorption current at constant electric field is independent of the dielectric thickness [179]. If, however, there is an enhanced density of dipoles near the interfaces, which may result from contamination or oxidation conditions, the current varies inversely with thickness. The absorption currents in response to positive and negative charging voltages should be mirror images [179].

A dipolar polarization has been found to occur in phosphosilicate glass films obtained by exposing thermally grown SiO_2 to a phosphorus diffusion cycle [26, 188]. It is doubtful that an adequate concentration of polar groups exist in pure SiO_2 in sufficient quantities to cause the observed relaxations. As explained in Sec. 2.1.2, dipole relaxations occur too slowly to explain the experimental results at 1μs-1ms time scales relevant to most analog circuits.

Charge injection resulting in trapped space charge effects

Electrons which are injected into the insulator from the contacts during stress can become trapped in defects distributed throughout the insulator bulk, leading to a buildup of uncompensated space charge. The buildup of this space charge within the insulator modifies the potential and decreases the electric field near the cathode, leading to a decay of the injected current with time which can be shown to follow a power law of the form Eq. 5.1 [189]. Various models based on this space-charge limited current flow have been proposed [175, 171] and this mechanism is believed to be responsible for absorption currents in polymers [178]. The field, temperature and contact dependence depend on the specific injection mechanism at the barrier which may occur by Schottky emission [176, 190, 180], Fowler-Nordheim tunneling or power law injection [178].

When a capacitor which has been charged at a high voltage for a prolonged period is suddenly short-circuited, electrons can be emitted from traps by a thermally activated process into the insulator conduction band from which they drift toward the electrodes. The potential in the insulator during the discharge period is controlled by the space-charge. Variations of this basic mechanism have been used to explain depolarization currents in thermally grown SiO_2 [181] and Ta_2O_5 [172] films. It has been suggested that discharge currents in TiO_2 are due to the emission of electrons to the insulator conduction band from traps localized near the interface with the positive electrode [173].

Space charge effects resulting from the injection of carriers into the oxide conduction band are assumed to be unimportant at the bias levels relevant to analog circuits where injection, and hence space-charge buildup, are negligible.

Electrode polarization

In imperfect dielectrics such as glass, conduction is largely due to the migration of ions in an electric field [162, 169, 191, 170]. Ionic drift may also occur in MIS insulators due to ionic contamination [21, 22, 188]. If the electrodes form blocking contacts, then this ionic motion causes a buildup of positive and negative charges at the negative and positive contacts respectively. The buildup of space charge at the contacts distorts the potential [189], resulting in a polarization current that decays with time. This electrode polarization effect has been used to explain absorption currents in electrolytic materials [162, 169, 192].

The absorption current due to electrode polarization depends linearly on the electric field and is thermally activated [169]. Clearly, the effect is strongly dependent on the electrode material, particularly the degree to which carriers are blocked or discharged [192] at the contacts. The thickness dependence is uncertain, but has been reported to be quadratic at constant field [178]. The charging and discharging currents are mirror images.

Electrode polarization is important primarily in glasses and other materials that behave as electrolytes. It can be assumed unimportant in uncontaminated device-grade oxides at time scales which are sufficiently short for ionic drift to be negligible.

Hopping between localized states

The hopping of a charge carrier between two localized potential wells has been shown to be indistinguishable from a dipole [193, 194, 195, 183]. In addition, it is possible for hopping conduction to occur between multiple localized states. If one assumes a random array of defect centers, this this leads to a wide distribution of relaxation times arising from the distribution of tunneling distances between trap sites. This model predicts that absorption currents should be linearly proportional to field and thermally activated [179]. In addition, the currents at constant field are independent of insulator thickness, independent of the electrode material and symmetrical for postive and negative voltages [179].

We shall assume that oxide traps are sufficiently far apart that tunneling transitions between them are improbable, and justify this assumption based on the data in Chap. 7.

Interfacial tunneling

A model for absorption currents due to the tunneling of electrons between metal electrodes and near-interface oxide traps was first proposed by Cherki et al [174]. Wilcox [196] later identified and corrected an error in this analysis and calculated absorption currents for the special case of traps distributed exponentially in energy and space. Wintle [197] extended the analysis to the case of a discrete trap energy

level and provided formulas for the absorption current and the dielectric constant and loss valid for low applied fields ($q\mathcal{E}x \ll kT$). The tunneling model was found to be in reasonable agreement with experimental results on Ta_2O_5 [174, 196] but was inconsistent with data on polymers (e.g. PET) [178]. The tunneling model has not yet been applied to insulators of importance to silicon integrated circuits, e.g. SiO_2 and Si_3N_4. In contrast to all of the other models discussed above, tunneling is not thermally activated, consistent with the very weak temperature dependence of our data.

A general tunneling model is derived in the following section, using the theory of Chapters 3 and 4. A new analysis is presented which focuses on the recovery voltage that develops across a capacitor as a function of time and voltage.

5.4 Trapping Theory

In this thesis, dielectric relaxation is measured by the following sequence, as shown in Fig. 1.3:

1. *Charge*; the capacitor is charged to a fixed voltage V_p for a period of time t_s.
2. *Discharge*; both plates are grounded and the capacitor is allowed to discharge for a period t_d.
3. *Float*; the top plate is open-circuited and the recovery voltage V_r that develops across the plates is monitored as a function of the relaxation time t_r, measured from the end of the discharge period.

Figure 5.2 shows energy band diagrams for a capacitor with metal plates and a distribution of traps in the dielectric near the interfaces. It is assumed that the insulator is sufficiently thick that tunneling transitions occur only between an oxide trap and the adjacent electrode. When a voltage $V_p > 0$ is applied to the top plate as shown in Fig. 5.2(b), empty traps at the bottom plate are lowered below the Fermi level where they can capture electrons while filled traps adjacent to the top plate rise above the Fermi level and emit electrons. When both capacitor plates are grounded, as shown in Fig. 5.2(c), filled traps rise above the Fermi level at the bottom plate and empty traps fall below the Fermi level at the top plate. If insufficient time is allowed for the trap occupancy to reach equilibrium during the discharge period, electrons will continue to flow out of traps to the bottom plate and into traps from the top plate during the float period [Fig. 5.2(d)]. In order to show how this tunneling action causes a recovery voltage to develop across the plates, we first consider an impulse of trapped charge $\Delta n_{ot}(E_t, x, 0)\Delta E_t \Delta x$ (cm^{-2}) in oxide traps near the interface with the bottom plate, with energies in a range ΔE_t around E_t and localized in depth between $x' = x$ and $x' = x + \Delta x$. The total response will then be calculated by superposition.

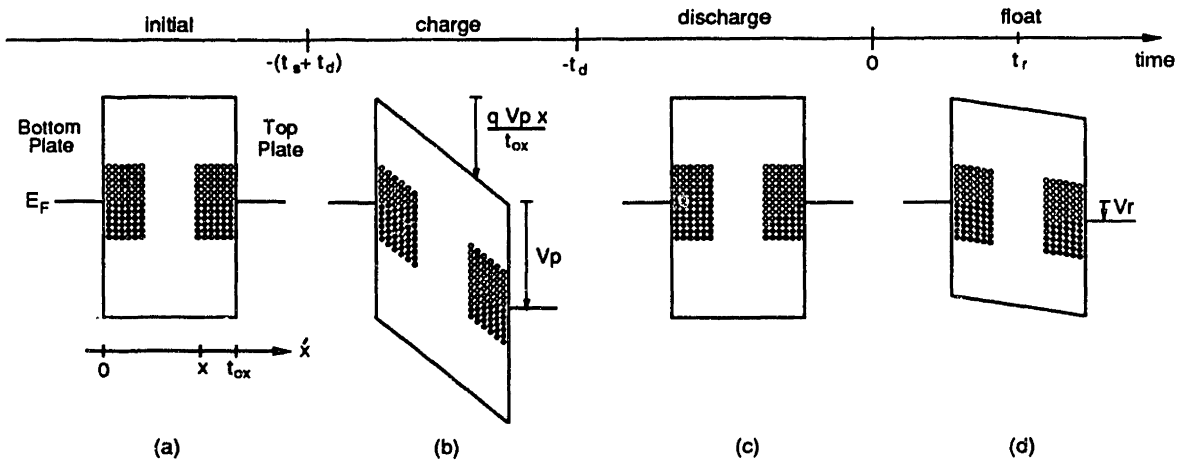


Figure 5.2: Energy band diagrams corresponding to dielectric relaxation measurement sequence: (a) initial (b) charge (c) discharge (d) float.

It was shown in Chap. 4 that the threshold voltage shift was independent of whether the traps were acceptor-like or donor-like. The same is true for the recovery voltage in capacitors, although the argument is not quite as obvious as Eq. ??. In the following section we assume acceptor-like traps, i.e. neutral when empty and negatively charged when occupied by an electron. Appendix F treats the case of donor-like traps and shows that the results are identical.

5.4.1 Trapping from the Bottom Plate

Positive Stress

When the charging voltage is positive, acceptor traps near the interface with the bottom plate fill with electrons during stress, becoming negatively charged. When the capacitor is short-circuited, these trapped electrons induce positive image charges on the plates in such a way as to make the total voltage drop across the

capacitor equal to zero. Most of this charge will be imaged on the bottom plate, which is closer, but some positive charge Q_r (Coul cm^{-2}) will be induced on the top plate as well. When the top plate is opened, this positive image charge, enclosed within dotted lines in Fig. 5.3(a)(i) is isolated and cannot change. Figs. 5.3(a)(ii)-(iv) show plots of the charge density, electric field and potential in the oxide at $t_r = 0^+$, immediately after the short-circuited capacitor is opened. Since the voltage across the capacitor cannot change instantaneously, the voltage across the capacitor is zero and the electric field lines, which have their source on positive image charges on the plates, all terminate internally on trapped electrons. As the trapped charge leaks out to the bottom plate by tunneling, electric field lines originating on the top plate begin to penetrate all the way through the dielectric and a voltage develops.

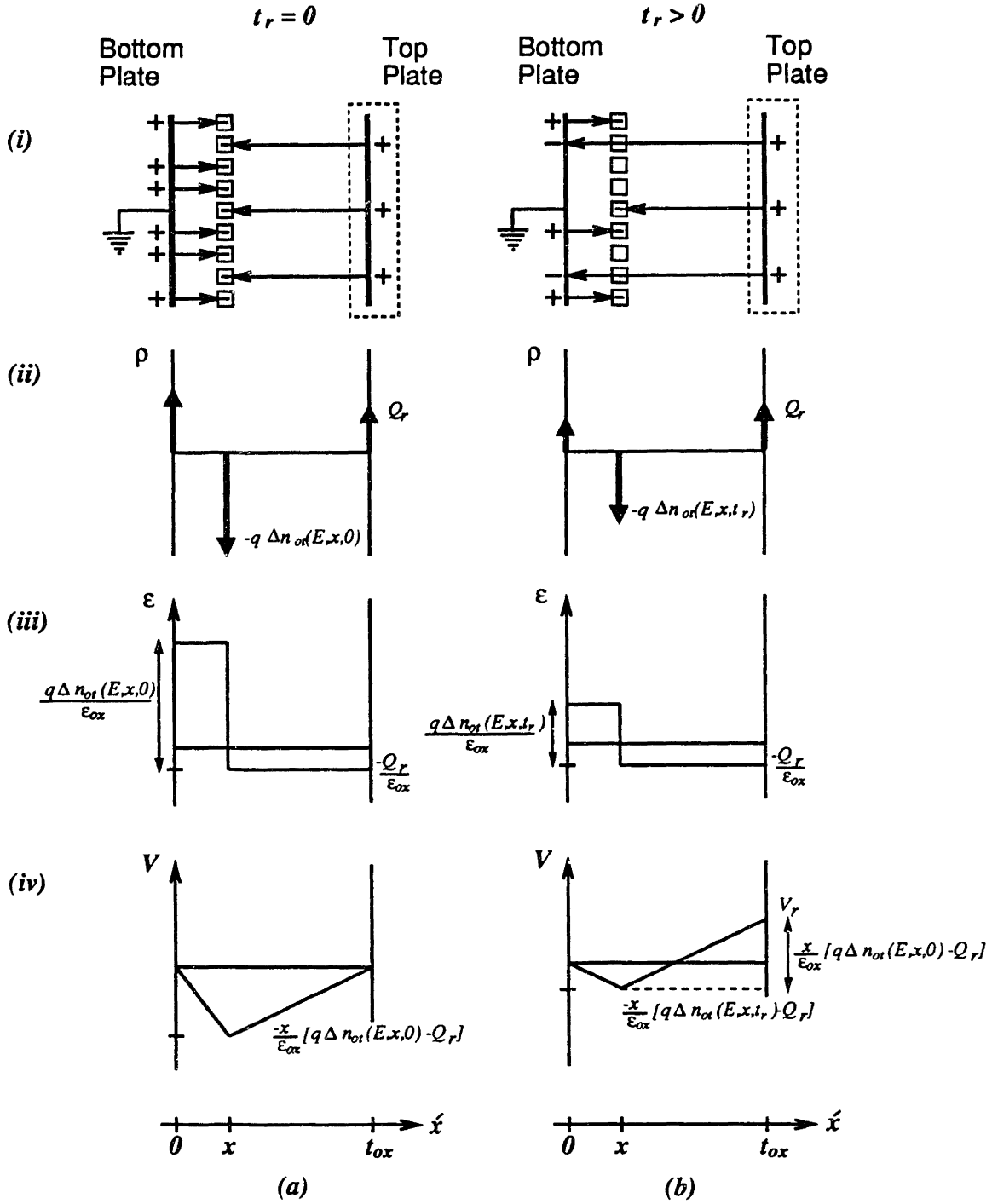


Figure 5.3: (i) Acceptor traps filled by tunneling from the bottom plate during positive stress, and plots of (ii) charge density (iii) electric field (iv) potential in the oxide. (a) Immediately following discharge period $[t_r = 0^+]$ (b) Float period $[t_r > 0]$.

The recovery voltage can easily be calculated with the aid of Fig. 5.3. The contribution $\delta\mathcal{E}(x', t_r = 0)$ to the electric field produced by the initial trapped charge distribution at depth $x' = x$ and energy E_t is, from Fig. 5.3(a)(iii),

$$\delta\mathcal{E}(x', 0) = \begin{cases} \frac{1}{\epsilon_{ox}}[q\Delta n_{ot}(E_t, x, 0)\Delta E_t \Delta x - Q_r] & 0 < x' < x \\ -\frac{Q_r}{\epsilon_{ox}} & x < x' < t_{ox} \end{cases} \tag{5.4}$$

Since the voltage across the capacitor cannot change instantaneously, $\delta V_r(t_r = 0^+) = \delta V_r(t_r = 0^-) = 0$, where $\delta V_r(0^+) = \delta V(t_{ox}, 0^+) - \delta V(0, 0^+)$. The variation of the potential in the oxide subject to this initial condition is therefore

$$\delta V(x', 0) = \begin{cases} -\frac{x'}{\epsilon_{ox}}[q\Delta n_{ot}(E_t, x, 0)\Delta E_t \Delta x - Q_r] & 0 < x' < x \\ -\frac{x}{\epsilon_{ox}}\left(\frac{t_{ox}-x'}{t_{ox}-x}\right)[q\Delta n_{ot}(E_t, x, 0)\Delta E_t \Delta x - Q_r] & x < x' < t_{ox} \end{cases} \tag{5.5}$$

Because the charge on the top plate is fixed, the electric field, and hence the slope of the potential, remain constant for $x < x' < t_{ox}$. However, as electrons tunnel out of oxide traps to the bottom plate, the electric field and hence the slope of the potential decrease for $0 < x' < x$ as shown in Fig. 5.3(b). More precisely, the electric field and potential vary with the relaxation time t_r as

$$\delta\mathcal{E}(x', t_r) = \begin{cases} \frac{1}{\epsilon_{ox}}[q\Delta n_{ot}(E_t, x, t_r)\Delta E_t \Delta x - Q_r] & 0 < x' < x \\ -\frac{Q_r}{\epsilon_{ox}} & x < x' < t_{ox} \end{cases} \tag{5.6}$$

$$\delta V(x', t_r) = \begin{cases} -\frac{x'}{\epsilon_{ox}}[q\Delta n_{ot}(E_t, x, t_r)\Delta E_t \Delta x - Q_r] & 0 < x' < x \\ -\frac{x}{\epsilon_{ox}}\{[q\Delta n_{ot}(E_t, x, t_r)\Delta E_t \Delta x - Q_r] & \\ \quad + \frac{x'-x}{\epsilon_{ox}}\left(\frac{x}{t_{ox}-x}\right)[q\Delta n_{ot}(E_t, x, 0)\Delta E_t \Delta x - Q_r]\} & x < x' < t_{ox} \end{cases} \tag{5.7}$$

where the second equation has used the fact that $Q_r = \frac{x}{t_{ox}} q\Delta n_{ot}(E_t, x, 0) dE_t dx$ or alternatively, by noting that the electric field for $x < x' < t_{ox}$ is the same in Figs. 5.3(a)(iii) and (b)(iii). A recovery voltage $\delta V_r(E_t, x, t_r) = \delta V(t_{ox}, t_r) - \delta V(0, t_r)$ develops across the plates as shown in Fig. 5.3(b)(iv). This voltage will increase as trapped charge tunnels out, eventually saturating when all the traps are empty. From Eq. 5.7,

$$\delta V_r(t_r) = \frac{qx}{\epsilon_{ox}}[\Delta n_{ot}(E_t, x, 0) - \Delta n_{ot}(E_t, x, t_r)]\Delta E_t \Delta x \tag{5.8}$$

The trapped electron concentration at the beginning of the float period, $\Delta n_{ot}(E_t, x, 0)$ is the result of charging the traps for a stress period of duration t_s and emission during a short-circuit period t_d. We can use the results derived for the cb/cb component in Chap. 4. None of the other components contribute because the plates are degenerate, with the Fermi level in the conduction band and effectively immobile. From Eq. 4.32,

$$\begin{aligned} \Delta n_{ot}(E_t, x, 0) = {} & D_{ot}(E_t, x)[f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x)] \\ & \times \{1 - \exp[-t_s/\tau_{ot/cb}(E_t + q\mathcal{E}_o x, x)]\} \exp[-t_d/\tau^o_{ot/cb}(E_t + q\mathcal{E}x, x)] \end{aligned} \tag{5.9}$$

with f and f_0 given by Eqs. 3.75 and 3.76. $\tau_{ot/cb}$ and $\tau^o_{ot/cb}$ are the time constants during stress and during emission, respectively, given by Eq. 3.18. This initial trapped charge distribution decays during the float period as

$$\Delta n_{ot}(E_t, x, t_r) = \Delta n_{ot}(E_t, x, 0) \exp[-t_r/\tau^o_{ot/cb}(E_t + q\mathcal{E}_o x)] \tag{5.10}$$

The time constant τ^o_{ot} during the float period is assumed equal to its value during the discharge period. Strictly speaking, this is not exactly true because the development of a recovery voltage causes a small difference between the electric field across the oxide during the discharge and float periods. However, since the recovery voltage is only a very small percentage of the original charging voltage V_p, and since the field dependence of the time constant is weak, this approximation results in negligible error.

By superposition, the total recovery voltage V^b_r due to trapping from the bottom plate, is obtained by substituting Eqs. 5.9 and 5.10 into Eq. 5.8 and integrating over energy and depth:

$$\begin{aligned} V^b_r(t_r) &= \int_0^{t_{ox}} \int_0^{\phi_c} \delta V_r(E_t, x, t_r) dE_t dx \\ &= \frac{q}{\epsilon_{ox}} \int_0^{t_{ox}} \int_0^{\phi_{cb}} x[f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x)] D_{ot}(E_t, x) \\ &\qquad \times [1 - \exp(-t_s/\tau_{ot/cb}(E_t + q\mathcal{E}x, x))] \exp(-t_d/\tau^o_{ot/cb}(E_t + q\mathcal{E}_o x, x)) \\ &\qquad \times \{1 - \exp[-t_r/\tau^o_{ot/cb}(E_t + q\mathcal{E}_o x, x)]\} dE_t dx \end{aligned} \tag{5.11}$$

where the integration extends from the bottom of the silicon conduction band to the SiO_2 conduction band. For $V_p > 0$, $f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x) > 0$ so that $V^b_r > 0$, as shown in Fig. 5.3(b)(iv).

Negative Stress

Traps rise relative to the Fermi level at the interface during stress with a negative voltage and emit electrons, becoming neutral. During the short-circuit period these traps are lowered below the Fermi level and begin to fill with electrons. Provided that t_d is shorter than the longest trap time constant, some of the traps remain empty at the beginning of the float period ($t_r = 0$) as shown in Fig. 5.4(a)(i).

Since they are initially neutral, the electric field and potential due to these traps are zero at $t_r = 0$ and there is no charge on the plates. As traps fill during the float period, an electric field builds up in the dielectric. Since the charge on the top plate is isolated, it remains zero and the field is therefore non-zero only for $0 < x' < x$, as shown in Fig. 5.4(b)(iii). It is clear from the Fig. 5.4(b)(iv) that the contribution of these traps to the recovery voltage is

$$\delta V^b_r = -\frac{qx}{\epsilon_{ox}} \Delta n_{ot}(E_t, x, t_r) \Delta E_t \Delta x \tag{5.12}$$

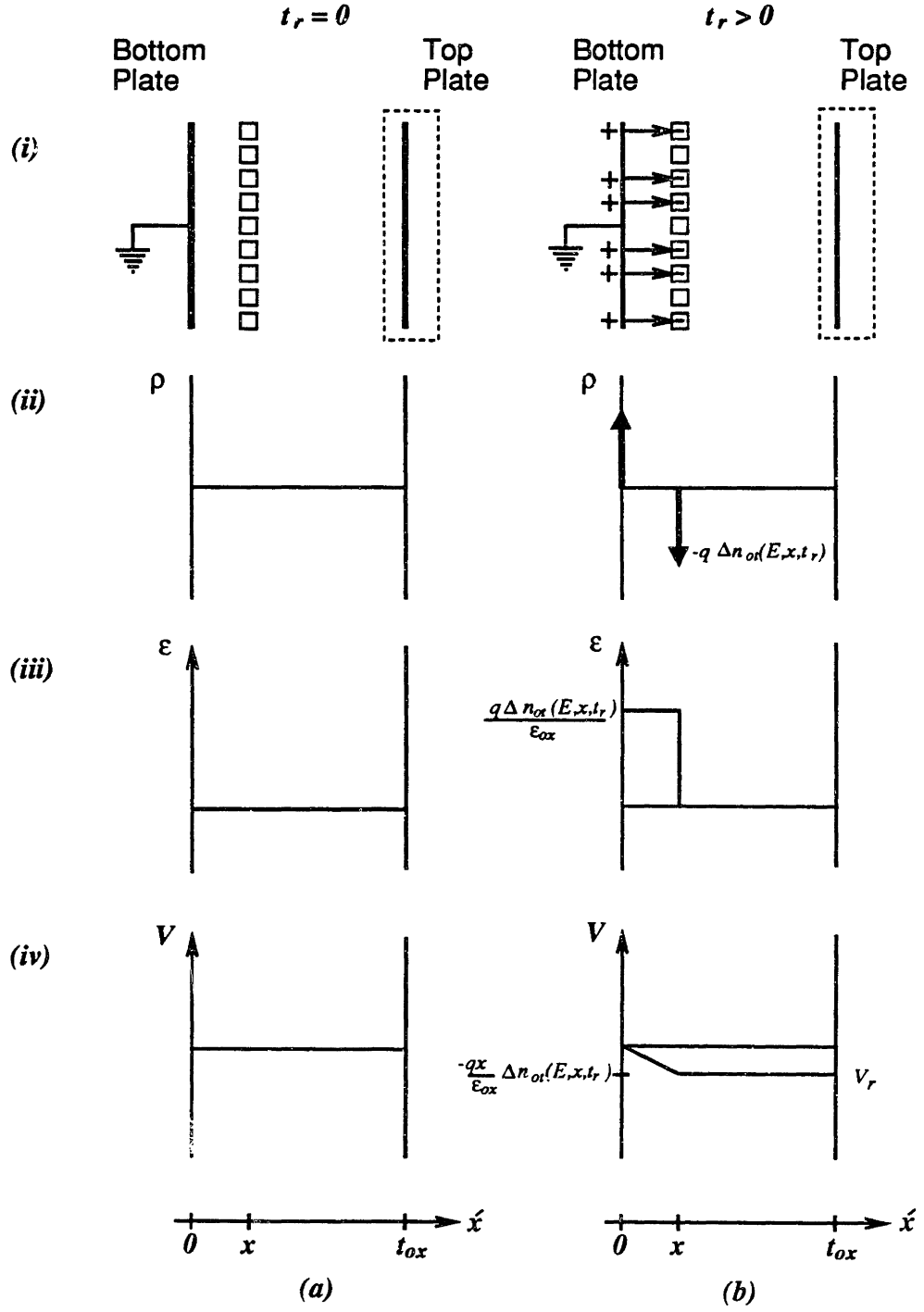


Figure 5.4: (i) Acceptor traps emptied by tunneling to the bottom plate during negative stress, and plots of (ii) charge density (iii) electric field (iv) potential in the oxide. (a) Immediately following discharge period $[t_r = 0^+]$ (b) Float period $[t_r > 0]$.

The filling of the traps is governed by

$$\Delta n_{ot}(E_t, x, t_r) = -\Delta n_{ot}(E_t, x, 0)\{1 - \exp[-t_r/\tau^o_{ot/cb}(E_t + q\mathcal{E}_o x, x)]\} \tag{5.13}$$

where $-\Delta n_{ot}(E_t, x, 0) = \Delta p_{ot}(E_t, x, 0)$ is the density of empty traps at $t_r = 0$ and $\Delta n_{ot}(E_t, x, 0)$ is the change in the density of filled traps which occured during the stress and discharge periods, given by Eq. 5.9.

Equation 5.13 simply states that as $t_r \to \infty$, all of the traps which were emptied during the stress period will be refilled. Substituting Eq. 5.9 in Eq. 5.13,

$$\begin{aligned}\Delta n_{ot}(E_t, x, t_r) = -D_{ot}(E_t, x)[f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x)] \\ \times \{1 - \exp[-t_s/\tau_{ot/cb}(E_t + q\mathcal{E}x, x)]\} \exp[-t_d/\tau^o_{ot/cb}(E_t + q\mathcal{E}_o x, x)] \\ \times \{1 - \exp[-t_r/\tau^o_{ot/cb}(E_t + q\mathcal{E}_o x, x)]\}\end{aligned} \tag{5.14}$$

From Eqs. 5.12, 5.13 and 5.14 it is clear that the recovery voltage is again given by Eq. 5.11 and is negative since $f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x) < 0$ for $V_p < 0$.

In Appendix F it is shown that Eq. 5.11 also holds for donor-like traps. Equation 5.11 is therefore a general expression for the recovery voltage due to charge trapping from the bottom plate and can be integrated numerically for an arbitrary trap distribution.

It is apparent from Eq. 5.11 that, if the time constants during and after stress are equal ($\tau_{ot/cb} = \tau^o_{ot/cb}$) and the trap distribution is uniform around E_F, the recovery voltage will be symmetrical with respect to postive and negative charging voltages. More will be said about this issue later.

5.4.2 Trapping from the Top Plate

We now consider the effects of tunnel exchange with the top plate. In order to emphasize the symmetry of the problem, we redefine the x axis in this section with its origin at the top plate, as shown in Fig. 5.5

Positive Stress

When subjected to a positive charging voltage, traps at the top plate rise relative to the Fermi level at the interface and emit electrons. These empty traps capture electrons during the discharge and float periods, becoming negatively charged as shown in Fig. 5.5(b)(i). The analysis parallels that of Fig. 5.4 and the contribution to the recovery voltage is easily seen from Fig. 5.4(b)(iv) to be

$$\delta V^t_r = \frac{q\Delta n_{ot}(E_t, x, t_r)\Delta E_t \Delta x}{\epsilon_{ox}} x \tag{5.15}$$

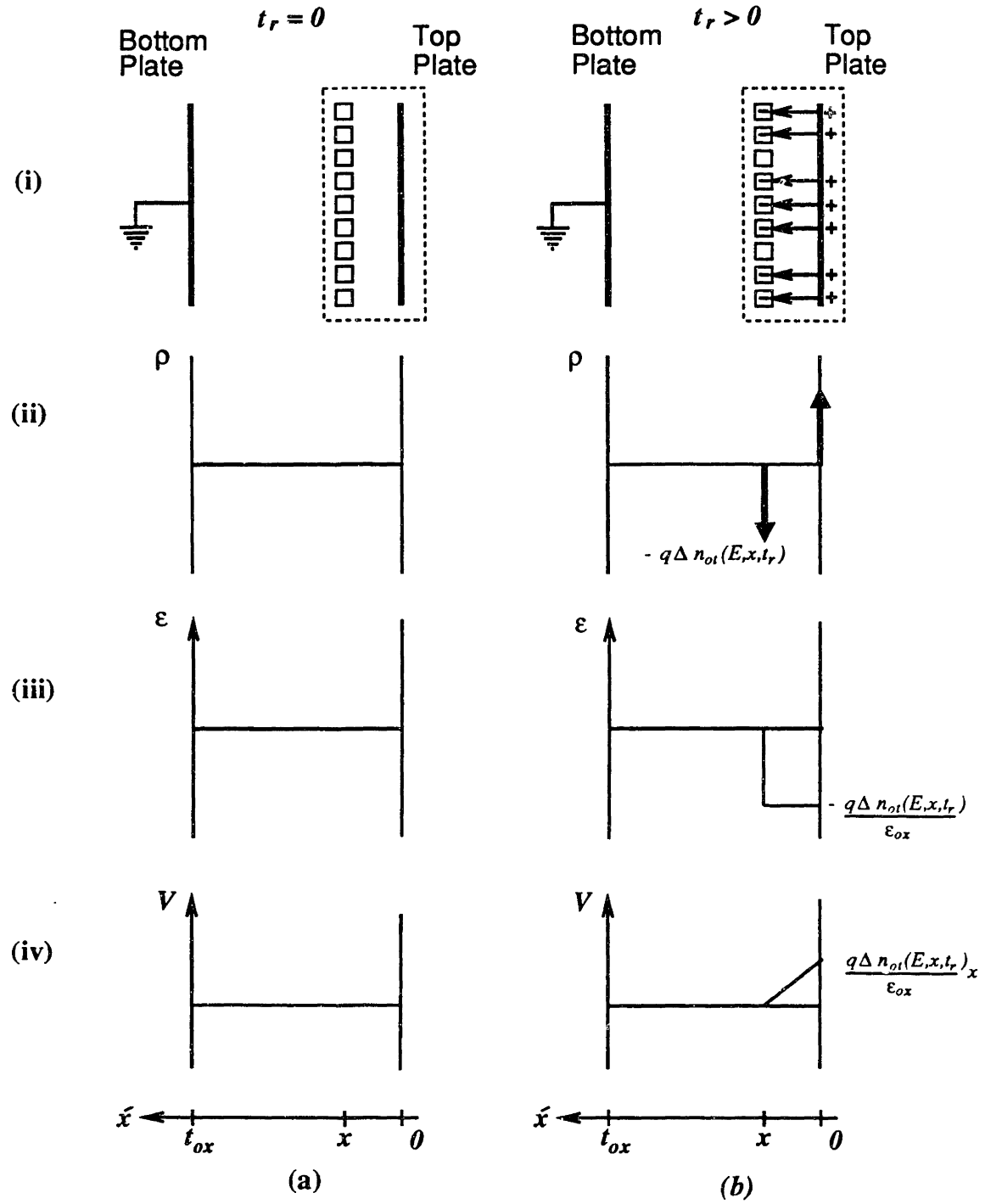


Figure 5.5: (i) Acceptor traps emptied by tunneling to the top plate during positive stress, and plots of (ii) charge density (iii) electric field (iv) potential in the oxide. (a) Immediately following discharge period $[t_r = 0^+]$ (b) Float period $[t_r > 0]$.

with $\Delta n_{ot}(E_t, x, t_r)$ given by Eq. 5.14. Integrating over energy and depth, the recovery voltage has the same form as Eq. 5.11 except for a minus sign,

$$V_r^t(t_r) = -\frac{q}{\epsilon_{ox}} \int_0^{t_{ox}} \int_0^{\phi_c} x[f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x)] D_{ot}(E_t, x) \times \{1 - \exp[-t_s/\tau_{ot/cb}(E_t + q\mathcal{E}x, x)]\} \exp[-t_d/\tau^o_{ot/cb}(E_t + q\mathcal{E}_o x, x)] \times \{1 - \exp[-t_r/\tau^o_{ot/cb}(E_t + q\mathcal{E}_o x, x)]\} dE_t dx \quad (5.16)$$

Since $f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x) < 0$ for $V_p > 0$, the recovery voltage V_r^t is positive as apparent from Fig. 5.5(b)(iv).

Negative Stress

For a negative charging voltage, traps adjacent to the top plate fill during stress, becoming negatively charged, as shown in Fig. 5.6(a)(i). During the discharge and float periods, they emit electrons, becoming neutral as shown in Fig. 5.6(b)(i). The charge enclosed by the dashed lines in Fig. 5.6 is isolated when the plates are open-circuited so that the field for $x < x' < t_{ox}$ is fixed. The contribution to the recovery voltage is, from Fig. 5.6(b)(iv),

$$\delta V_r^t = -\frac{-qx}{\epsilon_{ox}} [\Delta n_{ot}(E_t, x, 0) - \Delta n_{ot}(E_t, x, t_r)] \Delta E_t \Delta x \quad (5.17)$$

which is just the negative of Eq. 5.8. Substituting for $\Delta n_{ot}(E_t, x, 0)$ and $\Delta n_{ot}(E_t, x, t_r)$ from Eqs. 5.9 and 5.10 and integrating, the result is again given by Eq. 5.16. Since $f(E_t q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x) > 0$ for $V_p < 0$, $V_r < 0$, consistent with Fig. 5.6(b)(iv). Thus, Eq. 5.16 represents the recovery voltage due to trapping from the top plate, for either polarity of charging voltage. Appendix F demonstrates that the result holds regardless of whether the traps are acceptor- or donor-like.

The total recovery voltage is obtained by summing Eqs. 5.11 and 5.16,

$$V_r(t_r) = V_r^b(t_r) + V_r^t(t_r) \quad (5.18)$$

It has been shown above that both contributions are positive for $V_p > 0$ while both are negative for $V_p < 0$. Thus, the contributions from the bottom and the top plate superpose constructively and recovery voltage is the same polarity as the charging voltage.

5.5 Model

The recovery voltage can be calculated exactly by evaluating Eq. 5.18 with Eqs. 5.11 and 5.16. In this section we seek a simplified model subject to the following approximations

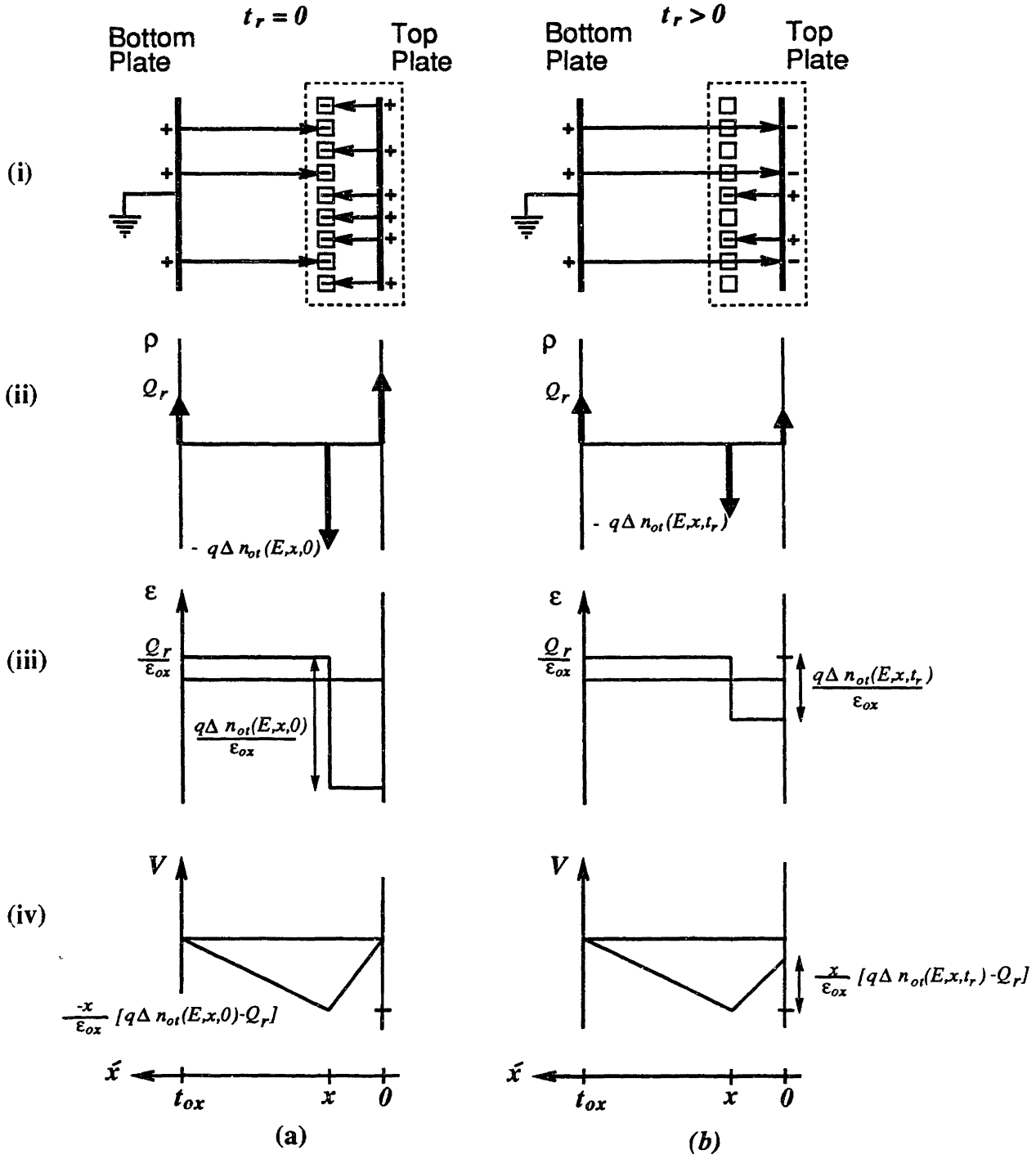


Figure 5.6: (i) Acceptor traps filled by tunneling from the top plate plate during negative stress, and plots of (ii) charge density (iii) electric field (iv) potential in the oxide. (a) Immediately following discharge period $[t_r = 0^+]$ (b) Float period $[t_r > 0]$.

1. Uniform trap density, $D_{ot}(E_t, x) = D_{ot}(\mathrm{cm}^{-3}\ \mathrm{eV}^{-1})$

2. Rectangular tunneling barrier with a time constant independent of energy and electric field

$$\tau_{ot/cb}(E_t, x) = \tau_{ot}(x) = \tau_{oc}\exp(x/x_o) \tag{5.19}$$
$$\tau^o_{ot/cb}(E_t, x) = \tau^o_{ot}(x) = \tau_{oe}\exp(x/x_o) \tag{5.20}$$

   where $x_o = \hbar/2\sqrt{2m^*(\phi_c - E)}$. We continue to allow for differences in the time constant prefactors τ_{oc} and τ_{oe} during and after stress.

3. Step approximation to the Fermi functions [Eq. 4.53].

   In this case, since we are dealing with metal electrodes, the step function approximation should be better than in the case of a MOSFET because the Fermi level is in a band where the density of states is large.

   Assuming $E_{F0} = E_F$ for degenerate plates, the third assumption implies

$$f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x) = \begin{cases} -1 & -q\mathcal{E}x < E_t < q\mathcal{E}_o x \\ 1 & -q\mathcal{E}_o x < E_t < -q\mathcal{E}x \\ 0 & \text{otherwise} \end{cases} \tag{5.21}$$

For tunneling from the bottom plate, the first inequality holds only for $V_p < 0$ and the second holds only for $V_p > 0$. For tunneling from the top plate, the first inequality holds only for $V_p > 0$ and the second holds only for $V_p < 0$.

With these approximations, Eq. 5.11 becomes, for either polarity of V_p,

$$V_r^b(t_r) \approx \frac{q^2 D_{ot} V_p}{\epsilon_{ox} t_{ox}} \int_0^{t_{ox}} x^2 [1 - e^{-t_s/\tau_{ot}(x)}] e^{-t_d/\tau^o_{ot}(x)} [1 - e^{-t_r/\tau^o_{ot}(x)}] dx \tag{5.22}$$

$$= \frac{q^2 D_{ot} V_p}{\epsilon_{ox} t_{ox}} \int_0^{t_{ox}} x^2 \left\{ [1 - e^{-t_s/\tau_{ot}(x)}] e^{-t_d/\tau^o_{ot}(x)} - [1 - e^{-t_s/\tau_{ot}(x)}] e^{-(t_d+t_r)/\tau^o_{ot}(x)} \right\} dx \tag{5.23}$$

where $|V_p| = (|\mathcal{E}| - |\mathcal{E}_o|)t_{ox}$ is the voltage applied across the oxide during stress. With degenerate plates, all of the voltage is dropped across the oxide so that the recovery voltage is directly proportional to the change in voltage. Applying Eqs. 5.21 to Eq. 5.16, the recovery voltage due to the top plate V_r^t is given by an equation of the same form as Eq. 5.23. The trap densities and time constant prefactors may, however, differ for the two plates if they are different materials.

As in Chap. 4 we begin with a simple analysis to gain intuitive understanding and then proceed with a more rigorous evaluation of the integral in Eq. 5.22.

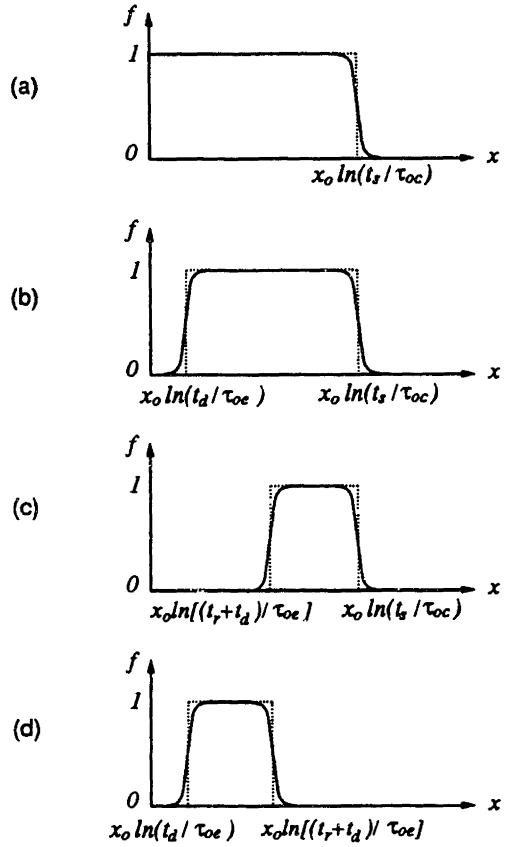


Figure 5.7: Spatial distribution of trapped charge during the dielectric absorption measurement sequence. (a) $t = t_s$ (b) $t = t_s + t_d$ (c) $t = t_s + t_d + t_r$ (d) traps which contribute to the relaxation integral for $t_r > 0$.

5.5.1 Simple Analysis

In this section we undertake an approximate solution for the recovery voltage in order to gain intuition into the general features of the voltage and time dependence. Recall that one factor of x in Eq. 5.23 arises from the fact that, unlike the threshold voltage shifts, the recovery voltage is due to image charges induced on the opposite plate from which the trapping occurs. This image charge increases proportional to the depth x of the trapped charges from the interface. The second factor of x arises, as in the it/cb component of the threshold shifts, from band bending in the oxide which causes the energy range of traps that are lowered below the Fermi level during stress to increase proportional to the depth of the traps from the interface.

Figure 5.7 shows the distribution of trapped charge in the oxide at various times during the dielectric absorption measurement sequence, using the approximation of Eqs. 4.58 and 4.59. In this approximation, traps fill during stress to a maximum

depth of $x_o \ln(t_s/\tau_{oc})$ as shown in Fig. 5.7(a). After a discharge time t_d, this filled trap distribution empties to a depth of approximately $x_o \ln(t_d/\tau_{oe})$, as shown in Fig. 5.7(b). It has been assumed that $t_s \gg t_d$, as is usually the case in practice, so that a rectangular approximation to the distribution is valid. Thus, the first term in the integrand of Eq. 5.23, which is represented in Fig. 5.7(b), is approximately unity over the range $x_o \ln(t_d/\tau_{oe}) < x < x_o \ln(t_s/\tau_{oc})$ and zero elsewhere. The recovery voltage is due to the decay of this initial charge distribution during the relaxation time t_r, as described in Sec. 5.4. After a time $0 < t_r < t_s - t_d$, the second term in Eq. 5.23, represented by Fig. 5.7(c), is approximately unity for $x_o \ln[(t_r + t_d)/\tau_{oe}] < x < x_o \ln(t_s/\tau_{oc})$. The recovery voltage is proportional to the difference between the first term [Fig. 5.7(b)] and the second term [Fig. 5.7(c)], which is approximately unity for $x_o \ln(t_d/\tau_{oe}) < x < x_o \ln[(t_r + t_d)/\tau_{oe}]$. When $t_r + t_d \approx t_s$ all the filled traps in Fig. 5.7(c) have emptied so that the integral is unity for $x_o \ln(t_d/\tau_{oe}) < x < x_o \ln(t_s/\tau_{oc})$ and the recovery voltage saturates. With these considerations, Eq. 5.23 can be approximated as

$$V_r(t_r) \approx \frac{q^2 D_{ot} V_p}{\epsilon_{ox} t_{ox}} \int_{x_L}^{x_H} x^2 dx$$

where

$$\begin{aligned} x_L &= x_o \ln\left(\frac{t_d}{\tau_{oe}}\right) \\ x_H &= \min\left\{x_o \ln\left(\frac{t_r + t_d}{\tau_{oe}}\right), x_o \ln\left(\frac{t_s}{\tau_{oc}}\right)\right\} \end{aligned}$$

Thus,

$$V_r(t_r) \approx \begin{cases} \frac{q^2 D_{ot} V_p x_o^3}{3\epsilon_{ox} t_{ox}} \left[\ln^3\left(\frac{t_r+t_d}{\tau_{oe}}\right) - \ln^3\left(\frac{t_d}{\tau_{oe}}\right)\right] & t_r \ll t_s - t_d \\ \frac{q^2 D_{ot} V_p x_o^3}{3\epsilon_{ox} t_{ox}} \left[\ln^3\left(\frac{t_s}{\tau_{oc}}\right) - \ln^3\left(\frac{t_d}{\tau_{oe}}\right)\right] & t_r \gg t_s - t_d \end{cases} \tag{5.24}$$

The saturation value of the recovery voltage increases with stress time as $\ln^3(t_s)$ and decreases with discharge time as $\ln^3(t_d)$. The absorption current obtained from the above expression is, for $t_d \ll t_r \ll t_s$,

$$\begin{aligned} i &= C_{ox} \frac{dV_r}{dt_r} \\ &= \frac{q^2 D_{ot} V_p {x_o}^3}{t_{ox}^2} \frac{1}{t_r} \ln^2\left(\frac{t_r}{\tau_{oe}}\right) \end{aligned}$$

In the case of low fields ($\mathcal{E} < 1 \times 10^5$ V/cm) this agrees with the result of Wilcox [196] obtained by a much more complicated derivation.

5.5.2 Exact Analysis

We now integrate each of the terms in the integrand of Eq. 5.22 exactly.

$$V_r(t_r) = \frac{q^2 D_{ot} V_p}{\epsilon_{ox} t_{ox}} \int_0^{t_{ox}} x^2 [e^{-t_d/\tau_{ot}^o(x)} - e^{-(t_s/\tau_{ot}(x)+t_d/\tau_{ot}^o(x))} - e^{-(t_d+t_r)/\tau_{ot}^o(x)} + e^{-[t_s/\tau_{ot}(x)+(t_d+t_r)/\tau_{ot}^o(x)]}] dx \tag{5.25}$$

We require the integral,

$$I = \int_0^{t_{ox}} x^2 e^{-t/\tau(x)} dx \tag{5.26}$$

with

$$\tau(x) = \tau_o \exp(x/x_o)$$

In order to simplify the integration, we first evaluate the integral

$$-\frac{dI}{dt} = \int_0^{t_{ox}} x^2 e^{-t/\tau} \frac{dx}{\tau}$$

where $\tau = \tau_o \exp(x/x_o)$. Changing the variable of integration from x to τ,

$$-\frac{dI}{dt} = x_o^3 \int_{\tau_o}^{\tau(t_{ox})} \ln^2 \frac{\tau}{\tau_o} e^{-t/\tau} \frac{d\tau}{\tau^2} \tag{5.27}$$

With the substitution $u = t/\tau$, Eq. 5.27 becomes

$$-\frac{dI}{dt} = \frac{x_o^3}{t} \left\{ \ln^2 \left(\frac{t}{\tau_o}\right) \int_0^\infty e^{-u} du - 2 \ln \left(\frac{t}{\tau_o}\right) \int_0^\infty \ln u e^{-u} du + \int_0^\infty \ln^2 u e^{-u} du \right\} \tag{5.28}$$

Generally, $\tau_o \approx 10^{-10}$ second and $\tau(t_{ox}) \sim 10^{26}$ years for a 100$\mathring{A}$ oxide, so that for the time scales of interest, the limits of integration can, to excellent approximation, be replaced by $t/\tau_o = \infty$ and $t/\tau(t_{ox}) = 0$. The second integral in Eq. 5.28 is a standard form [198]

$$\int_0^\infty \ln u e^{-u} du = \Gamma'(1) = -\gamma$$

where Γ is the gamma function and $\gamma = -0.57721$ is Euler's constant. The third integral can be bounded by separating the integration limits into two intervals [196]

$$\int_0^\infty e^{-u} \ln^2 u du < \int_0^1 \ln^2 u du + \int_1^\infty u^2 e^{-u} du$$

since $e^{-u} < 1$ for $0 < u < 1$ and $u > \ln u$ for $1 < u < \infty$. Using [199, 200]

$$\int_0^1 \ln^2 u du = 2$$

$$\int_1^\infty u^2 e^{-u} du = \frac{5}{e} = \frac{5}{2.718}$$

Equation 5.28 can be written,

$$-\frac{dI}{dt} = \frac{x_o^2}{t}\left\{\ln^2\left(\frac{t}{\tau_o}\right) + 1.154\ln\left(\frac{t}{\tau_o}\right) + \delta\right\}$$

where $\delta < 3.84$. Since $\tau_o \sim 10^{-10}$s, for the time scales of interest, ($t > 10^{-6}$s), the $\ln^2(t/\tau_o)$ term dominates and

$$-\frac{dI}{dt} \approx \frac{x_o^3}{t}\ln^2\left(\frac{t}{\tau_o}\right)$$

The original integral, Eq. 5.26 is then

$$\int_0^{t_{ox}} x^2 e^{-t/\tau} dx \approx -\frac{x_o^3}{3}\ln^3\left(\frac{t}{\tau_o}\right) + C \tag{5.29}$$

where C is a constant of integration. Substituting Eq. 5.29 in Eq. 5.25 and using Eqs. 5.19 and 5.20, the recovery voltage is

$$\begin{aligned} V_r(t) \;=\; & \frac{q^2 D_{ot} x_o^3 V_p}{3\epsilon_{ox} t_{ox}}\left\{\ln^3\left(\frac{t_d + t_r}{\tau_{oe}}\right) - \ln^3\left(\frac{t_s}{\tau_{oc}} + \frac{t_d + t_r}{\tau_{oe}}\right)\right. \\ & \left. + \ln^3\left(\frac{t_s}{\tau_{oc}} + \frac{t_d}{\tau_{oe}}\right) - \ln^3\left(\frac{t_d}{\tau_{oe}}\right)\right\} \end{aligned} \tag{5.30}$$

This equation is the main result of this chapter. It is well behaved at $t_r = 0$, where $V_r(t_r = 0) = 0$, even though the analysis is strictly valid only for $t_r \gg \tau_o$. At very long times, the recovery voltage saturates at

$$V_r(t_r \to \infty) = \ln^3\left(\frac{t_s}{\tau_{oc}} + \frac{t_d}{\tau_{oe}}\right) - \ln^3\left(\frac{t_d}{\tau_{oe}}\right) \tag{5.31}$$

It is easily verified that Eq. 5.30 reduces to the approximate result, Eq. 5.24, in the appropriate limits.

5.6 Summary

In this chapter it has been shown that:

- The recovery voltage is given by the sum of contributions from the bottom plate, given by Eq. 5.11 and from the top plate, given by Eq. 5.16. These contributions superpose constructively. This result is valid for both acceptor-like and donor-like oxide traps.
- An approximate expression for the recovery voltage, in the case of a uniform trap distribution, is given by Eq. 5.30. This equation applies equally well to the top and bottom plates. The recovery voltage is directly proportional to the charging voltage V_p and varies as $\ln^3$(time).

- For a symmetrical capacitor structure, the recovery voltage for positive and negative stress voltages will always be symmetrical and of the same polarity as the charging voltage. In the case of an asymmetrical capacitor structure, for example one in which trapping occurs from only one plate, it is clear from Sec. 5.4 that the positive and negative recovery voltages will still be symmetrical provided that the trap distribution is uniform and the time constants during capture and emission are equal. However, if trap densities are nonuniform so that the change in trap occupancies during positive and negative charging voltages differ or their filling and emptying rates differ, then asymmetrical behavior will result.

- It is clear from Eq. 5.30 that, for constant electric field $\mathcal{E} = V_p/t_{ox}$, the recovery voltage is independent of the oxide thickness. Thus, the relative error V_r/V_p increases with decreasing t_{ox} at constant $\mathcal{E}$. This differs from the trapping-induced threshold voltage shifts, which scale with t_{ox} so that the relative error is independent of oxide thickness at constant electric field.

- The temperature dependence is contained in the Fermi functions in the integrands of Eqs. 5.11 and 5.16. As in the case of threshold voltage relaxation, the temperature dependence will depend strongly on the energy distribution of oxide traps. Since the Fermi distribution spreads towards higher energies with increasing temperature, it is expected that the recovery voltage will be relatively independent of temperature for a uniform D_{ot}, increase with temperature for a D_{ot} which increases with energy, and decrease with temperature if D_{ot} decreases with energy.

In Chapter 7, simulation results, using the models derived in this chapter, will be presented and compared with experimental data.

Chapter 6

Measurement Techniques

6.1 Introduction

Measurements of relaxation effects in MOS transistors and capacitors must achieve two objectives:

1. The emulation, as closely as possible, of the operating environment of typical analog circuit applications.

2. The isolation of charge trapping effects from contributions due to other sources.

The first goal is important because, ultimately, we are interested in the impact of relaxation effects on circuit performance. Many device studies are conducted in bias regions outside the ranges of interest to circuit operation and there is a danger in extrapolating these results to devices in circuits. For example, DLTS studies on capacitors with p-type substrates are normally performed by pulsing the device from deep depletion into accumulation and then back, thereby studying traps in the lower half of the bandgap [72, 84]. The NMOS transistor is of course never deep-depleted because of the presence of the source and drain contacts but more importantly, for operation in the inversion region, it is traps in the upper half rather than in the lower half of the bandgap which are important. Trapping studies on MOS capacitors are often performed with constant current stress which may create traps. Threshold shifts in MOSFETs are usually observed under high voltage bias conditions far exceeding actual circuit conditions [59] and may create damage in the oxide by hot carrier injection [41]. In many of these measurements it is unclear whether the threshold voltage or flatband voltage shift is due to thermal drifts, the creation of damage, or by filling and emptying of pre-existing traps. While these experiments are useful for studying the physics of charge trapping in insulators, they are not of much use to circuit designers, interested in the properties of MOS dielectrics under typical circuit conditions.

From a circuit perspective, the best way to measure these effects is in an actual circuit environment, e.g. a charge redistribution A/D converter. This approach was employed in a previous study [10], where threshold voltage relaxation was measured by differentially stressing the input transistors of a MOS amplifier and monitoring the long-tail settling component at the output. While this investigation demonstrated the importance of threshold voltage relaxation as a limiting factor on analog circuit performance, it was unable to conclusively isolate the source of the problem as being due solely to charge trapping. MOSFETs biased in analog circuits are typically biased in saturation where various other mechanisms, including thermal hysteresis and hot carrier effects, may also contribute to threshold voltage instabilities. Our present purpose is primarily to explain and model these results. Consequently, the second goal is necessary in order to determine whether the threshold voltage shifts are, in fact, due to interfacial tunneling based on the theory developed in the first five chapters.

Unfortunately, these two goals are not completely compatible. As discussed in the following section, the second objective requires that measurements be made in the linear region, which is not a normal bias condition for MOSFETs in analog circuits. Normal saturation region operation, as dictated by the first goal, may lead to additional contributions to the threshold voltage shifts. While these contributions are certainly of concern from a circuit point of view, they complicate the determination of the role of charge trapping, which is the topic of this thesis. These tradeoffs are discussed in Sec. 6.2. Section 6.3 then describes the measurement of dielectric relaxation in capacitors.

6.2 Threshold Voltage Relaxation

6.2.1 Introduction

A circuit designed to measure trapping-induced threshold voltage shifts in MOSFETs must achieve the following specifications:

1. Isolation of the device under test (DUT)
   The main challenge in the design of a measurement system is in the elimination of measurement artifacts and long-tail responses of circuit components such as other transistors, power supplies, resistors and opamps. Almost any circuit component will behave nonlinearly and exhibit relaxation effects under large-signal stress conditions. This goal dictates simplicity and as few extraneous circuit components as possible.

2. Gain
   The small amplitudes of the trapping-induced threshold voltage shifts (100μV-1mV) presents a measurement challenge at time scales (1μs-1ms) relevant to analog circuit applications. A gain of approximately 100 is adequate for most

measurements. Preferably, the gain should be independent of the DUT and bias conditions, so that it need not be remeasured at each data point.

3. Speed
   The measurement circuit must be fast enough to detect the initial threshold shift before it dies away. Since the duration of the long-tail decreases with the initial amplitude of the threshold voltage shift, the circuit speed also dictates the smallest shift which can be observed. Since the long tails persist for times up to hundreds of milliseconds, there is no rigid requirement on circuit speed but the highest possible bandwidth acheivable is desired in order to compare theory and experiment over many decades.

4. Noise
   Input-referred noise should be less than 10μV in order to observe small threshold voltage shifts.

5. Accurate control of bias point
   The bias point of the DUT should be accurately controllable with low drift, so that threshold voltage shifts can be monitored over a range of voltage levels in both the saturation and linear regions. This argues against measuring the device within an A/D converter or amplifier as in Ref. [10] and in favor of a simple single-device test circuit.

6. Elimination of thermal contributions
   Changes in power dissipation in the DUT during stress can cause self-heating resulting in changes in the threshold voltage which are indiscernible from the effects which we are trying to measure. The test circuit design must prevent such thermal contributions.

7. Elimination of hot carrier contributions
   The circuit design must also prevent the injection of hot carriers during stress. Hot carriers can become trapped in the insulator causing threshold voltage shifts, and can also create new damage as discussed in Sec. 2.1.6. This objective is most easily accomplished by operating the DUT in the linear region with low drain-source voltage.

8. Elimination of oxide trap generation
   In addition to hot carriers generated under high transverse electric fields, oxide traps may be generated under high vertical electric fields in excess of about 1.5 MV/cm [201]. This puts a limit on the maximum pulse voltages which can be applied.

Table 6.1 summarizes some of the mechanisms which may contribute to threshold voltage shifts in circuits and indicates the experimental and/or measurement conditions which allow these effects to be neglected.

The final goal and also the desire for low power dissipation recommend that the measurements be taken in the linear region. Linear region measurements are also desirable for analytical simplicity, since with the same level of inversion along the entire channel, a one-dimensional band model can be used. Unfortunately, this complicates somewhat the gain problem.

In the linear region, the current is given by

$$I_{DS} = \beta \left[(V_{GS} - V_T) - \frac{V_{DS}}{2} \right] V_{DS} \tag{6.1}$$

where $\beta = \mu C_{OX} W/L = K'W/L$ and it is assumed that $V_{DS}/2 \ll V_{GS} - V_T$. The transconductance is

$$g_m = \left. \frac{\partial I_{DS}}{\partial V_{GS}} \right|_{V_{DS}} = \beta V_{DS} \tag{6.2}$$

The gain is $A_v = g_m(R_L \| \frac{1}{g_{DS}})$ where R_L the external load resistor and the output conductance is

$$g_{DS} = \left. \frac{\partial I_{DS}}{\partial V_{DS}} \right|_{V_{GS}} = \beta(V_{GS} - V_T) = \frac{I_{DS}}{V_{DS}} \tag{6.3}$$

Generally, $\frac{1}{g_{DS}} < R_L$ in the linear region, in which case the gain is

$$A_v \leq \frac{V_{DS}}{V_{GS} - V_T} < 1 \tag{6.4}$$

For typical values, $R_L = 10k\Omega$, $I_{DS} = 100\mu$A, $V_{DS} = 100$mV and $K' = 40\mu\text{A/V}^2$,

$$\begin{aligned} g_{DS} &= 1/1k\Omega \\ g_m &= 80\mu A/V \\ A_v &\approx 0.1 \end{aligned} \tag{6.5}$$

Thus, the linear region gain is usually less than one and some external amplification must be provided. Due to gain-bandwidth tradeoffs, it is not possible to completely satisfy the speed and resolution requirements simultaneously. We discuss below two alternative measurement techniques which emphasize speed and gain respectively.

6.2.2 Measurement Circuit

Open Loop Configuration

Figure 6.2 shows a simple circuit to measure threshold voltage relaxation in MOSFETs. The gate of the DUT is subjected to a voltage pulse by connecting a CMOS switch between a voltage supply and ground. The output is taken differentially between the drains of the DUT and a second matched MOSFET which can be trimmed

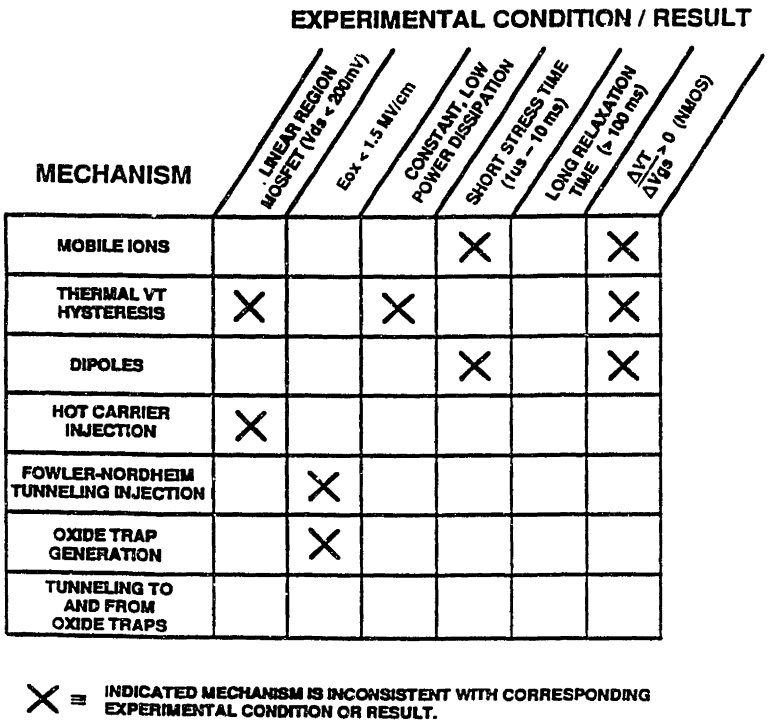


Figure 6.1: Mechanisms which may contribute to threshold voltage shifts and experimental conditions by which they are elimated.

so that the output is zero in the steady state with both gates grounded. The output is observed on a Tektronix 7854 oscilloscope with a 7A22 differential amplifier input. The input to the 7A22 is Schottky clamped to prevent overdriving the oscilloscope.

The sources are biased negatively with respect to the gates, which are normally at ground. Note that, since the sources are connected to a common voltage source rather than a current source, the connection is symettrical but not differential. This is important because, in a differential pair, one device may cutoff when the other is stressed with a positive voltage, and vice versa. In order to ensure that any observed threshold voltage shifts are due only to the DUT, it is important that the second device never be subjected to stress, which is accomplished by the symmetrical connection shown.

During stress, the gate of the DUT is switched to a voltage supply for a variable length of time. When the gate returns to ground, any shift in threshold voltage results in a deviation of the drain voltages which can be observed at the output. The output is averaged 1000 times and read into an HP series computer.

The primary advantage of the open loop circuit is speed. It suffers, however, from several drawbacks; (1) Low gain, as given by Eq. 6.4, (2) The gain of the device

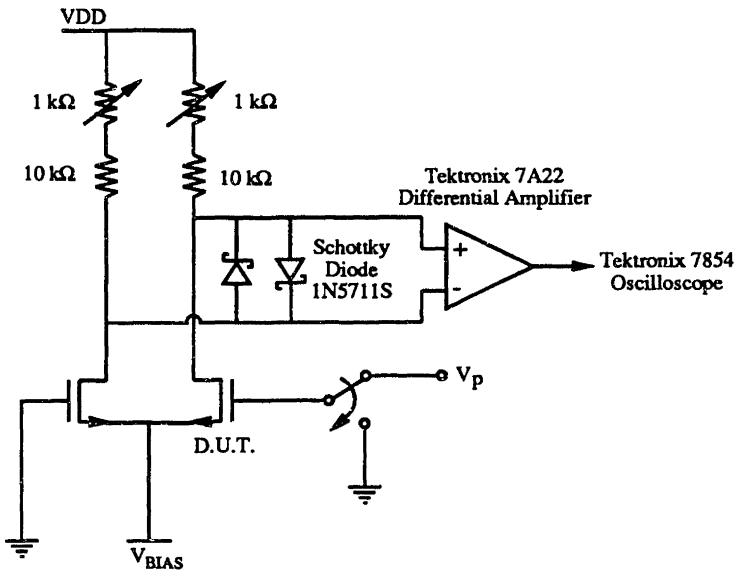


Figure 6.2: Open-loop measurement circuit.

under test must be remeasured for each device or bias condition in order calculate the input-referred value of the threshold shift, (2) the output of the DUT pulls down to the negative supply or up to the positive supply during positive and negative stress voltages, respectively. While the output is Schottky clamped at $\sim \pm 300$mV, there is still a danger of transient overdrive of the oscilloscope inputs since the scope gain must be large in order to overcome the linear region attenuation of Eq. 6.4. This circuit is therefore used only for measurements at short times. For all other measurements, particularly of threshold voltage shifts versus bias conditions, a closed-loop technique is used which overcomes these problems.

Closed Loop Configuration

The closed-loop measurement circuit, described in this section, directly measures the equivalent gate-source voltage shift, as described in Chap. 4. The DUT is connected in a loop with a high gain opamp in the feedback path, as shown in Fig. 6.3, where the CMOS switches are shown in their pre-stress positions. When the offset voltage of the differential pair is trimmed to zero, the loop forces the gate of the DUT to return to a well-defined reference level, in this case ground, after any threshold voltage shifts have died away.

During stress, the gate of the DUT is switched to a voltage source and the opamp is disconnected from the feedback path and put into a standby mode. This eliminates several effects which could interfere with the measurement: (1) hysteresis in the input stage of the opamp when the drain of the DUT pulls low, (2) saturation recovery time

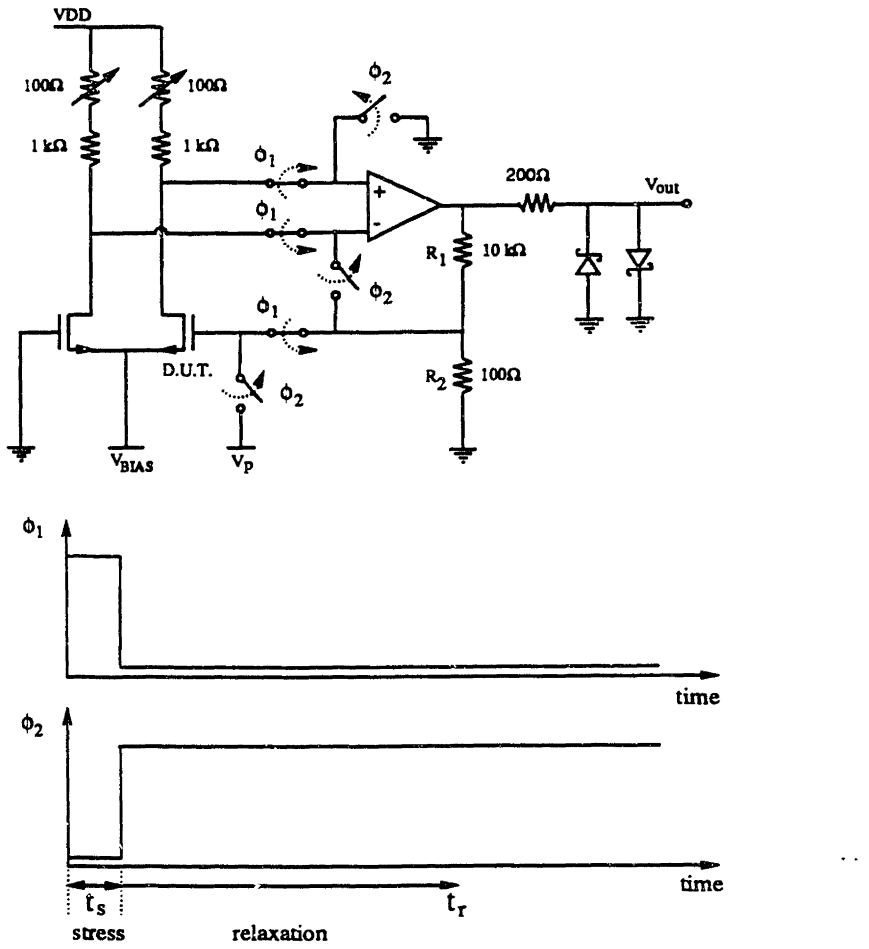


Figure 6.3: Closed-loop measurement circuit.

of the operational amplifier, and (3) thermal hysteresis due to differential heating of the input stage which may result from asymmetrical stressing of the input stage or by thermal feedback from the output stage when large currents are being supplied to the load. The offset voltage of the opamp is trimmed so that that V_{out} remains at ground during the stress period. This ensures that gate of the DUT is not stressed further when the feedback loop is reconnected.

After a 1μs to 10ms stress period, the switches are returned to their initial positions. Charge captured during stress changes the conductance of the channel due to small shifts in the threshold voltage and the mobility of inversion layer charge carriers. This tends to change the drain current of the DUT from its initial value. The feedback, however, acts to maintain a constant drain current by precisely compensating for any change in threshold voltage or mobility by a change in gate-source voltage. This *equivalent gate-source voltage* ΔV_{GS} is multiplied by an adjustable gain

factor

$$A_v = \frac{R_1 + R_2}{R_2} \tag{6.6}$$

and observed on an oscilloscope. The output is Schottky clamped in order to avoid overdriving the oscilloscope preamplifier which could also cause hysteresis effects. The measured transients are averaged 1000 times and acquired by an HP computer.

The closed-loop gain of the feedback loop, Eq. 6.6, is independent of the transconductance of the device being measured provided that the loop transmission is much greater than unity [202],

$$g_m \left(R_L \parallel \frac{1}{g_{DS}} \right) A_o \left(\frac{R_2}{R_1 + R_2} \right) \gg 1 \tag{6.7}$$

where A_o is the open loop gain of the opamp and R_L the load resistor. With the linear region g_m given by Eq. 6.2, the loop transmission approaches zero as V_{DS} goes to zero, so that Eq. 6.7 constrains the minimum value of V_{DS} and the maximum closed loop gain [Eq. 6.6] which can be used.

All components, including scope inputs, opamps, pulse generators and power supplies were thoroughly characterized to ensure that they did not contribute any long tail response, thermal or otherwise, to the response. Various opamps were tested and the AD744 and HA2542 were found to have sufficient speed and thermal recovery for the measurements. The diodes (1N4148) were also tested and found not to interfere with the circuit response. In most cases, it was found that the diodes could be removed from the circuit without affecting the response. High-speed CMOS switches (HI-201HS) with a 40ns switching time were more than adequate for the measurements.

6.2.3 Conditions for Constant Power Dissipation

For either of the measurement circuits discussed above, there is a unique combination of power supply voltage and drain resistor values such that the power dissipated in the DUT during and after stress is identical. The elimination of any differential power dissipation through the DUT ensures that thermal effects will not contribute to the measured threshold shift.

Figure 6.4 shows the bias points of the DUT before (V_{DS1},I_{DS1}) and during (V_{DS2},I_{DS2}) stress. The bias condition following stress is nearly identical to that before stress. Also shown is a curve of constant power dissipation $P_d = I_{DS}V_{DS}$=constant. We desire values for the external drain resistor R and power supply voltage V_{DD} such that the bias points before and during stress lie on the same constant power curve.

Before stress, the bias levels are specified by the desired initial level of inversion $V_{GS1} - V_T = \Delta V_1$ and drain-source voltage V_{DS1}. During stress, the desired amplitude of the stress voltage sets $V_{GS2} - V_T = \Delta V_2$ as shown in Fig. 6.4. The power dissipation

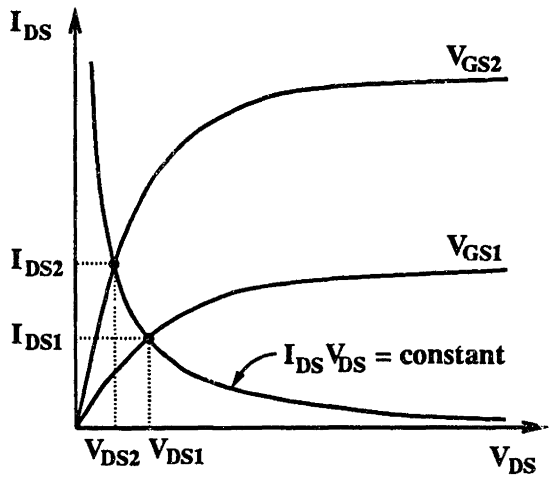


Figure 6.4: I-V characteristics before and after stress and constant power dissipation characteristic

is calculated from the steady-state currents and voltages using the static value of the threshold voltage V_T, with negligible error. The only remaining degree of freedom is the drain-source voltage during stress V_{DS2}, which can be adjusted by varying the drain resistor R and power supply voltage V_{DD} to make the power dissipation before, during and after stress equal.

The drain current before stress is

$$I_{DS1} = \beta\left(\Delta V_1 - \frac{V_{DS1}}{2}\right)V_{DS1} \tag{6.8}$$

$$= \frac{V_{DD} - V_{DS1}}{R} \tag{6.9}$$

where $\beta = \mu C_{OX} W/L$. During stress, the drain current is

$$I_{DS2} = \beta\left(\Delta V_2 - \frac{V_{DS2}}{2}\right)V_{DS2} \tag{6.10}$$

$$= \frac{V_{DD}}{R + R_{DS2}} \tag{6.11}$$

where

$$R_{DS2} = \left(\frac{\partial I_{DS2}}{\partial V_{DS2}}\right)^{-1} \approx \frac{1}{\beta \Delta V_2} \tag{6.12}$$

where linear region operation has been assumed, with $V_{DS2} << \Delta V_2$. The drain voltage during stress is then, from Eqs. 6.10 and 6.12,

$$V_{DS2} = I_{DS2} R_{DS2} = \frac{V_{DD}}{1 + \beta \Delta V_2 R} \tag{6.13}$$

The condition for constant power dissipation, $I_{DS1}V_{DS1} = I_{DS2}V_{DS2}$ requires that

$$\beta\left(\Delta V_1 - \frac{\beta V_{DS1}}{2}\right)V_{DS1}^2 = \frac{\Delta V_2 V_{DD}^2}{(1+\beta\Delta V_2 R)^2} \tag{6.14}$$

But,

$$\begin{aligned} V_{DD} &= I_{DS1}R + V_{DS1} \\ &= V_{DS1}\left[1 + \beta R(\Delta V_1 - V_{DS1}/2)\right] \end{aligned} \tag{6.15}$$

Substituting Eq. 6.15 into Eq. 6.14 and solving for R,

$$R = \frac{1 - \sqrt{(\Delta V_1 - V_{DS1}/2)/\Delta V_2}}{\beta\left[\Delta V_1 - \frac{V_{DS1}}{2} - \sqrt{\Delta V_2(\Delta V_1 - V_{DS1}/2)}\right]} \tag{6.16}$$

Thus, for a specified ΔV_1, ΔV_2 and V_{DS1}, Eq. 6.16 can be solved for R, which then determines V_{DD} from Eq. 6.15. These values of drain resistance and power supply voltage ensure that the power dissipation through the device under test does not change during the measurement sequence.

6.3 Dielectric Relaxation in Capacitors

Dielectric relaxation is measured using the circuit shown in Fig. 6.5. The circuit cycles the capacitor under test through the sequence depicted in Fig. 1.3 as described in Secs. 1.1.2 and 5.4.

During stress, the top plate of the capacitor under test is connected to ground through CMOS switch GT, and the bottom plate is connected to either the positive ($+V_p$) or negative ($-V_p$) reference voltages through switches PV or MV, respectively. During the discharge period, the capacitor bottom plate is switched to ground through switch GB. Note that, since the voltage across the capacitor cannot change instantaneously and because there is a finite resistance to ground through switch GT, a transient voltage spike may occur when switch GB is closed. In order to prevent transient voltage pulses from stressing the MOSFET M1 and causing threshold voltage relaxation effects to interfere with the measurement, the gate of the source-follower is grounded through switch R during the stress and discharge periods.

The relaxation period is initiated by opening switches GT and R and connecting the capacitor top plate to the gate of M1 through switch C. The output is taken differentially through the two source followers M1 and M2 and input to a Tektronix 7854 oscilloscope through a 7A22 differential amplifier. Each recovery voltage transient is averaged 1000 times and acquired by an HP computer for analysis.

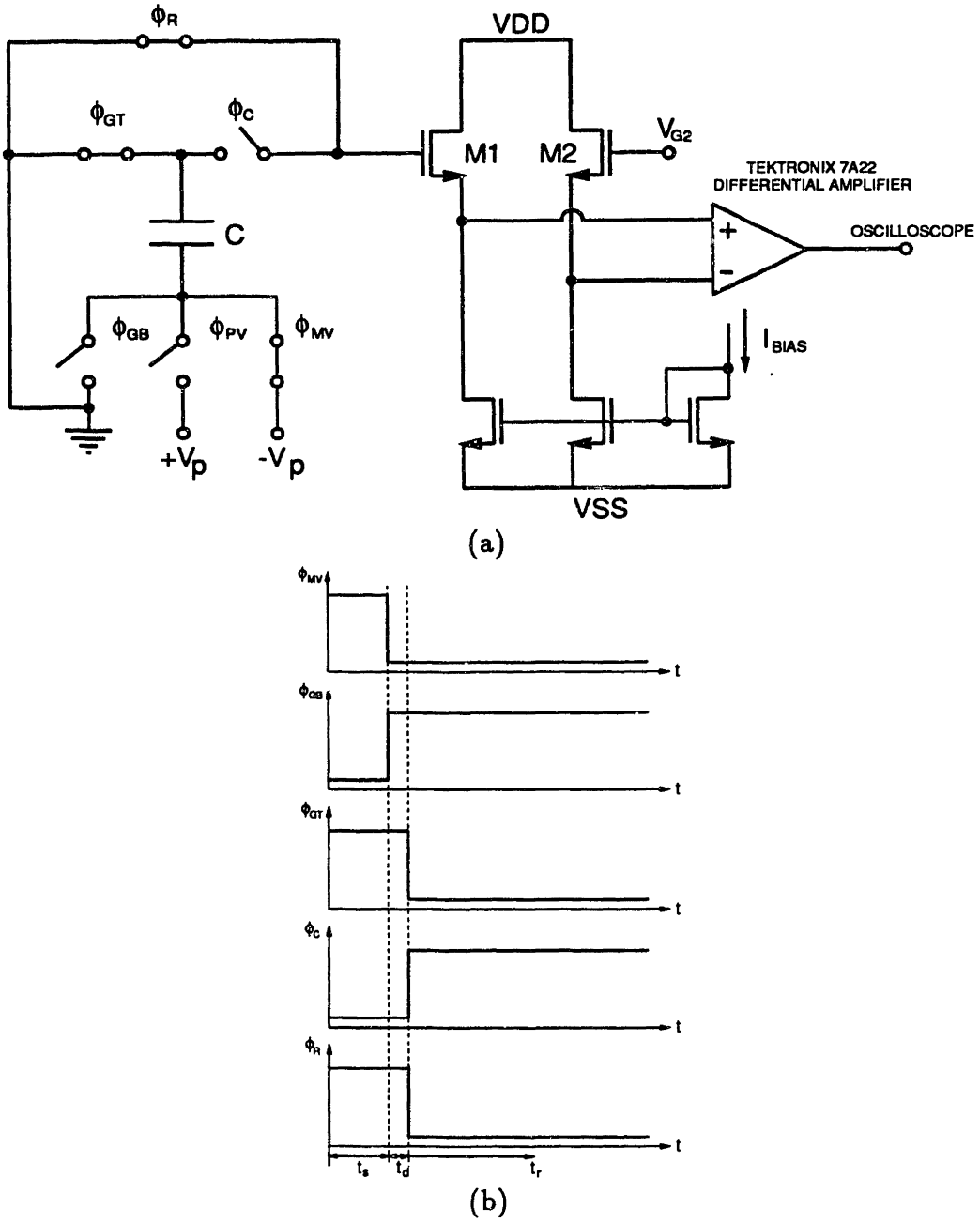


Figure 6.5: Dielectric relaxation measurement circuit; (a) simplified circuit schematic (b) timing diagram.

Chapter 7

Experimental and Simulation Results

7.1 Introduction

The first part of this chapter, Sec. 7.2, is devoted to experimental measurements of threshold voltage relaxation in MOSFETs. Data on conventional and nitrided SiO_2 dielectrics are presented as a function of bias conditions and explained qualitatively in terms of the theory of Chap. 4, with emphasis on the approximate models derived in Sec. 4.4. Quantitative simulation results are then presented in Sec. 7.2.4 for key experimental results, using the numerical models of Chaps. 3 and 4. The relationship between the measured threshold voltage shifts and 1/f noise in these processes is addressed in Sec. 7.2.5 and a strong correlation is demonstrated. Section 7.2.6 covers issues relevant to circuits, including a simple model for circuit simulation purposes, a discussion of the effects of threshold voltage relaxation on circuit accuracy and techniques to minimize these errors.

Dielectric relaxation in capacitors is the subject of Sec. 7.3. Experimental results are presented for silicon dioxide and nitrided oxide dielectrics. This data is compared with simulated recovery voltages using the models of Chapter 5.

7.2 Threshold Voltage Relaxation

Threshold voltage relaxation was characterized in eight different processes which are summarized in Table 7.1. In this section we present measurements from the first five processes, which use conventional thermally grown SiO_2 gate dielectrics. Devices from processes 1-4 were fabricated at commercial foundries while those from process 5 were fabricated at MIT's Microsystems Technology Laboratories. Threshold voltage relaxation in nitrided and reoxidized nitrided oxides (processes 6–8) are presented in Sec. 7.2.3 and compared with conventional SiO_2 dielectrics.

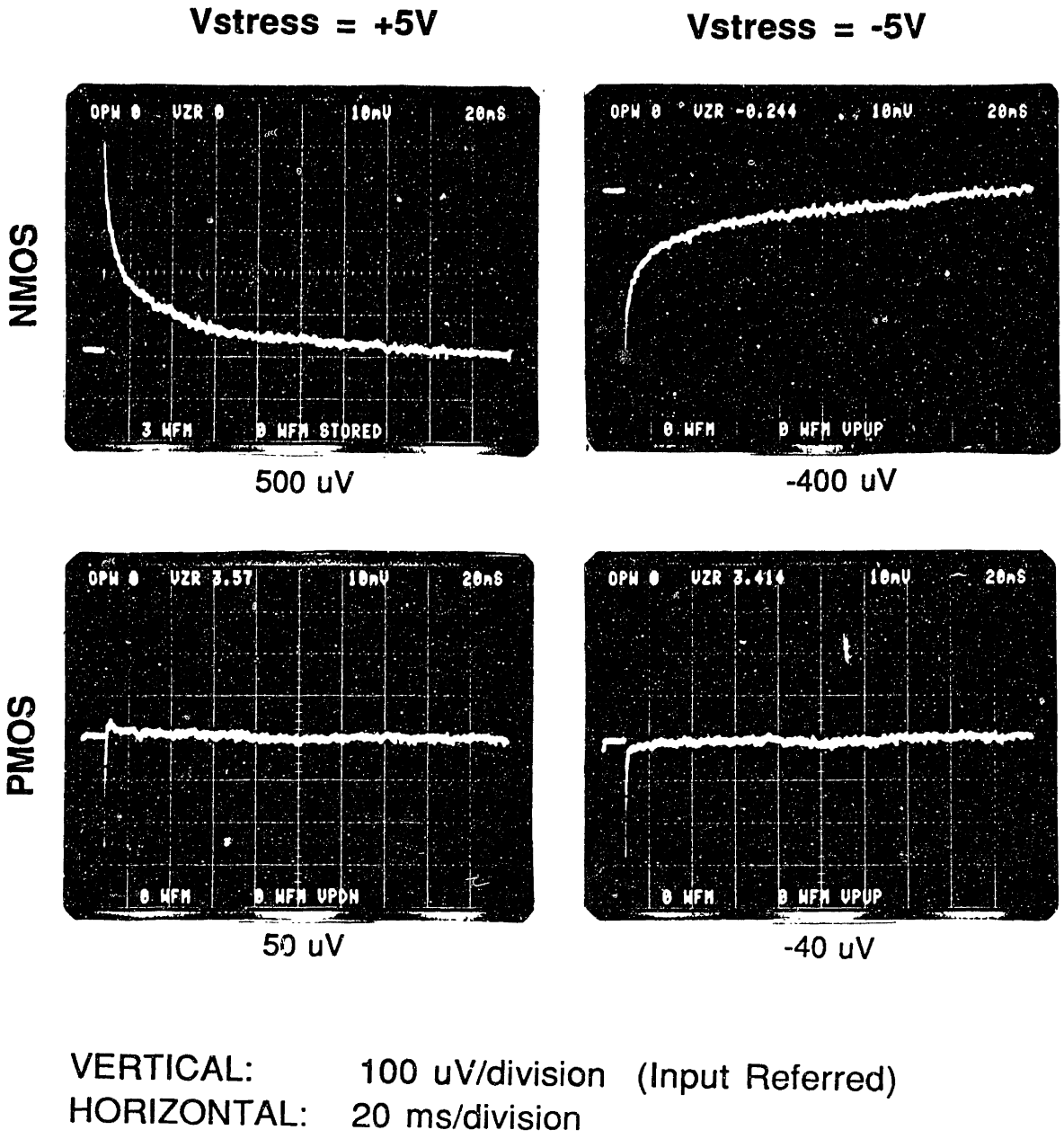


Figure 7.1: Oscilloscope photographs showing measured V_T shifts in process 5 (W/L = 20μm/5μm); t_s = 10ms, V_p = $\pm$5V, V_{DS}=350mV. Vertical scale = 100 μV per division (input-referred), Horizontal scale = 20ms per division.

Table 7.1: Process Summary

Process	Foundry	Dielectric	Thickness	W/L	NMOS V_T (V)	PMOS V_T (V)
1	Commercial	SiO_2	350Å	20/5	0.63	-0.92
2	Commercial	SiO_2	350Å	100/5	0.73	-0.87
3	Commercial	SiO_2	750Å	48/4	0.75	-0.76
4	Commercial	SiO_2	500Å	20/5	0.70	-1.5
5	MIT	SiO_2	250Å	20/5	0.70	-0.64
6	MIT	SiO_2	100Å	20/5	0.36	-1.28
7	MIT	NOX	100Å	20/5	0.30	-0.76
8	MIT	ROXNOX	100Å	20/5	0.25	-1.34

Threshold voltage relaxation transients observed at the output of the circuit of Fig. 6.3 are shown in the oscilloscope photographs of Fig. 7.1 for NMOS and PMOS transistors biased in the linear region and stressed with 10ms gate voltage pulses of amplitude ±5V. The vertical scale is 10mV per division, or 100uV/divison when referred to the input throught the circuit gain The NMOS transistor shows an input referred threshold voltage shift of 500uV when stressed with +5V, which decays with a very long tail that persists for more than 200ms. Negative stress also produces a threshold voltage shift, but the amplitude is smaller. The figure shows that threshold voltage shifts in PMOS devices are less than in NMOS by about a factor of ten. Some of the PMOS devices tested showed no discernible threshold voltage shift at the resolution of the measurement circuit ($\approx 10\mu$V) and hence can be used to calibrate the system. This gives us confidence that the response of the circuit is clean and does not contribute any slow settling components which would interfere with the measurement.

Tables 7.2 and 7.3 summarize measurements of linear region threshold voltage shifts observed in conventional SiO_2 gate dielectrics. The single number used as a figure of merit in these measurements is the mean amplitude of the "initial" threshold voltage shift $\overline{\Delta V_T}$ a short delay time t_d from the end of the stress period. The mean is taken, for each process, over 10-20 devices from one lot. Unless otherwise stated, the measurements in this chapter assume a stress time of t_s = 10ms, and the V_T shift is sampled a time t_d = 1ms after the falling edge of the stress voltage pulse. Note the rather large standard deviations of the data which tend to support a trapping mechanism in which the threshold voltage shifts depend on a random distribution of defects which may vary widely from device to device.

7.2.1 Thermal Considerations

Figure 7.2(a) shows the temperature dependence of the drain current of an NMOS transistor in process 2, measured using the open-loop circuit of Fig. 6.2 with fixed

Table 7.2: NMOS (t_s = 10ms, t_d = 1ms)

				$V_p = +5V$		$V_p = -5V$	
Process	W/L	$V_{GS}(V)$	$V_{DS}(mV)$	ΔV_T (μV)	σ (μV)	ΔV_T (μV)	σ (μV)
1	20/5	1.0	200	370.0	50.0	-135.0	22.4
2	100/5	1.0	200	700.1	113.5	-368.6	189.1
3	48/4	1.2	300	482.8	19.4	-135.0	44.6
4	20/5	1.0	200	65.2	28.0	-33.8	21.8
5	20/5	1.2	300	523.6	72.2	-432.8	236.0

Table 7.3: PMOS (t_s = 10ms, t_d = 1ms)

				$V_p = +5V$		$V_p = -5V$	
Process	W/L	$V_{GS}(V)$	$V_{DS}(mV)$	ΔV_T (μV)	σ (μV)	ΔV_T (μV)	σ (μV)
1	20/5	-1.5	500	27.2	26.1	-5.3	3.0
2	100/5	-1.5	500	37.3	28.1	-20.0	11.9
3	48/4	-2.0	500	37.2	10.0	-49.6	29.5
4	20/5	-2.0	500	11.9	2.3	-4.8	1.8
5	20/5	-1.4	300	134.6	93.4	-28.7	17.3

V_{GS} = 1.2V. The temperature coefficient of the drain current is seen to be about -0.2μA/°C. Since the feedback in the closed-loop measurement circuit (Fig. 6.3) forces constant current through the DUT, any self-heating which occurs during stress will result in a compensatory increase in gate voltage which will indiscernible from a threshold voltage shift at the output of the circuit.

In order to completely eliminate any vestige of thermal self-heating from our measurements, the experiments of Fig. 7.1 were conducted under conditions of constant power dissipation in the DUT. For example, in the NMOS measurement with $V_p = 5V$ we have $R_L = 1k\Omega$ and $VDD = -1.15V$ which yields a constant power dissipation level throughout the measurement sequence:

Before and after stress:	$I_{DS1} = 120\mu A$	$V_{DS1} = 450mV$	$P_{d1} = 54\mu W$
During stress:	$I_{DS2} = 270\mu A$	$V_{DS2} = 200mV$	$P_{d2} = 54\mu W$

For stress voltages of the opposite polarity the DUT cuts off and constant power dissipation cannot be maintained. However the measurements are carried out in the linear region with V_{DS} = 200mV so that device self-heating effects are minimized. In the following section we show that the drain voltage dependence of the threshold voltage shifts is sufficiently weak that constant power dissipation need not be strictly enforced in order to ensure that the observed response is purely trapping-related.

Unless otherwise stated, the closed-loop circuit of Fig. 6.3 was used for the measurements in this section. The open-loop circuit [Fig. 6.2] was used when greater speed was required, as well as for calibration purposes and to provide independent

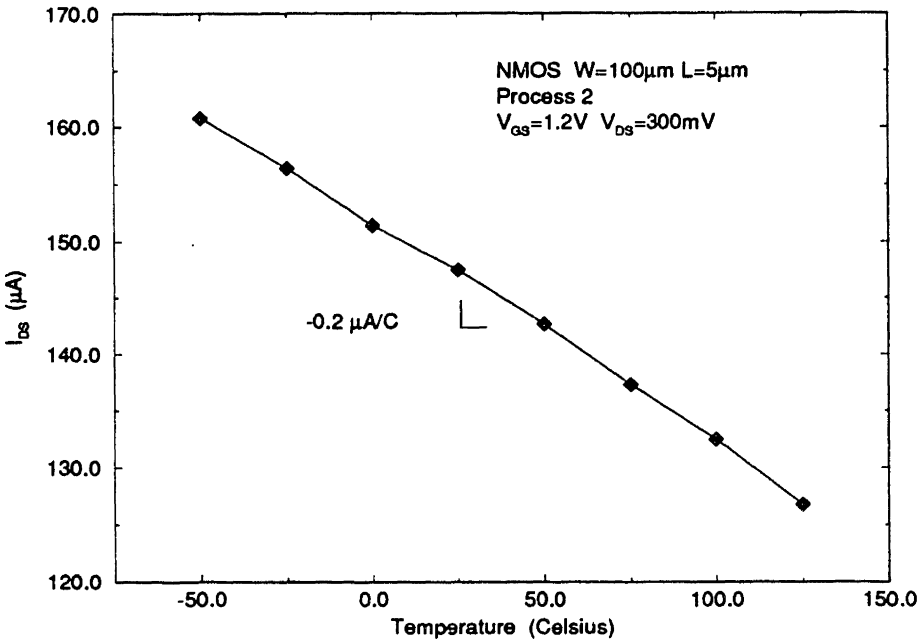


Figure 7.2: Drain current vs. temperature (process 2)

verification of the measurements from the closed-loop circuit. In all cases, both circuits were found to provide consistent measurements.

7.2.2 Dependence on Device Characteristics and Test Conditions

The bias dependence of the threshold voltage relaxations provides the key to understanding the underlying physics and provides the vehicle for verifying the theory presented in Chapters 3 and 4. Each process leaves its signature in the form of its dependence on voltage, time and temperature and from this we can learn many things about the nature of the mechanism and the distribution of oxide traps in energy and space. In this section we present experimental results for the dependence of the threshold voltage shifts on drain voltage, relaxation and stress times, stress voltage, initial gate bias, substrate bias, temperature and device geometry.

Drain Voltage

We now abandon the requirement for constant power dissipation and observe the magnitude of the threshold voltage shift as the pre-stress value of the drain-source voltage is varied. As a consequence of the variation of V_{DS}, the power dissipation before and during stress are no longer equal. Depending on bias conditions, the power dissipa-

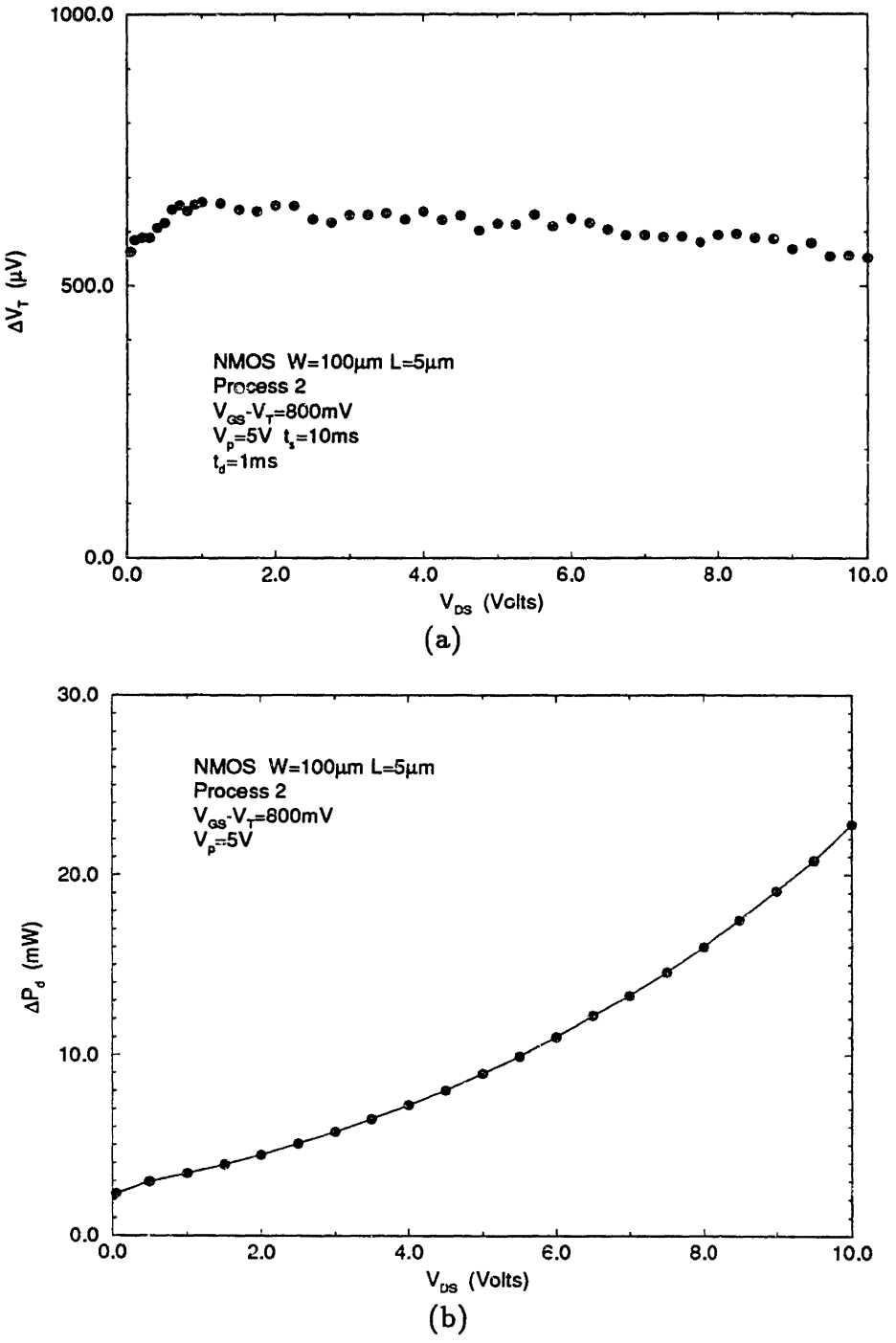


Figure 7.3: (a) Dependence of V_T shift on drain voltage, (b) change in power dissipation during stress as a function of drain voltage.

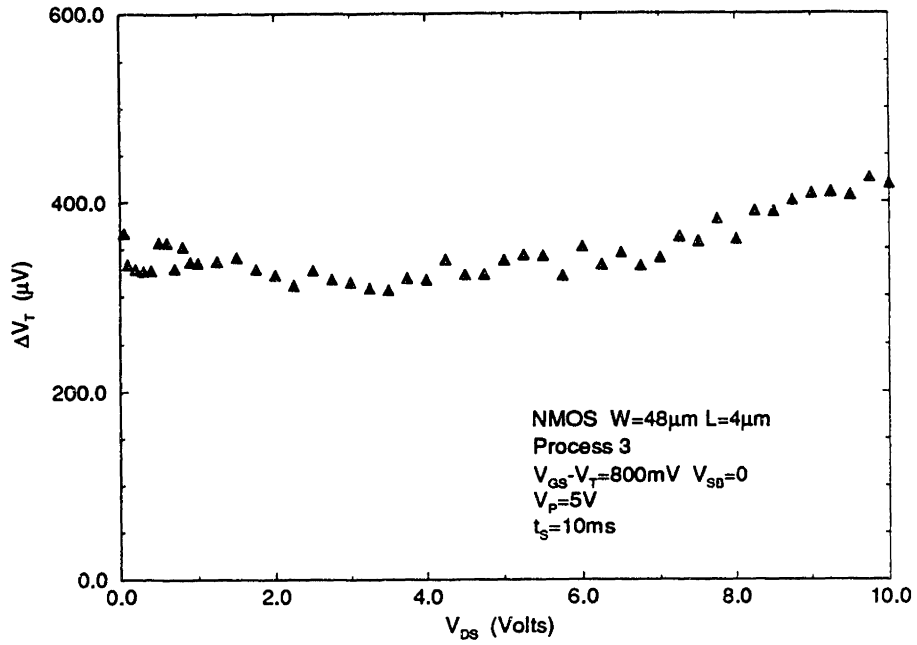


Figure 7.4: Drain voltage dependence of initial V_T shift in process 3.

tion in the DUT may either increase or decrease during stress. With $R_L = 10k\Omega$, $V_{GS} - V_T = 0.8V$ and $V_p = 5$V, the change ΔP_d in power dissipation which occurs during stress is positive and increases with the initial V_{DS} level as shown in Fig. 7.3(b). Any self-heating which occurs due to the increase in power dissipation through the DUT during stress will cause a momentary increase in gate voltage when the stress is removed, which will decay to zero as the device returns to thermal equilibrium. The measured gate voltage shift is actually very nearly constant, in fact decreasing slightly with drain voltage as shown in Fig. 7.3(a). We can conclude that thermal as well as hot carrier effects, both of which would would increase with increasing drain voltage, are unimportant at moderate drain voltages, even in saturation.

The reason for the slight decrease in ΔV_T with drain voltage at very low V_{DS} is due to the degradation of the transconductance (Eq. 6.2) and hence the loop transmission as discussed in Sec. 6.2.2. The slight decrease with increasing V_{DS} is not completely understood, and is not always observed. Another process [Fig. 7.4] shows a slight increase in ΔV_T with increasing V_{DS}, which may be due to hot carrier generation or self-heating at high V_{DS}. In either case, these effects can be eliminated by operating in the linear region with 200mV< V_{DS} <500mV where the change in power dissipation ΔP_d is small. In the remainder of this section we shall therefore relax our insistence on constant power dissipation, conducting all measurements in the linear region where it has been established that thermal contributions to the observed threshold voltage shifts are negligible.

Relaxation Time

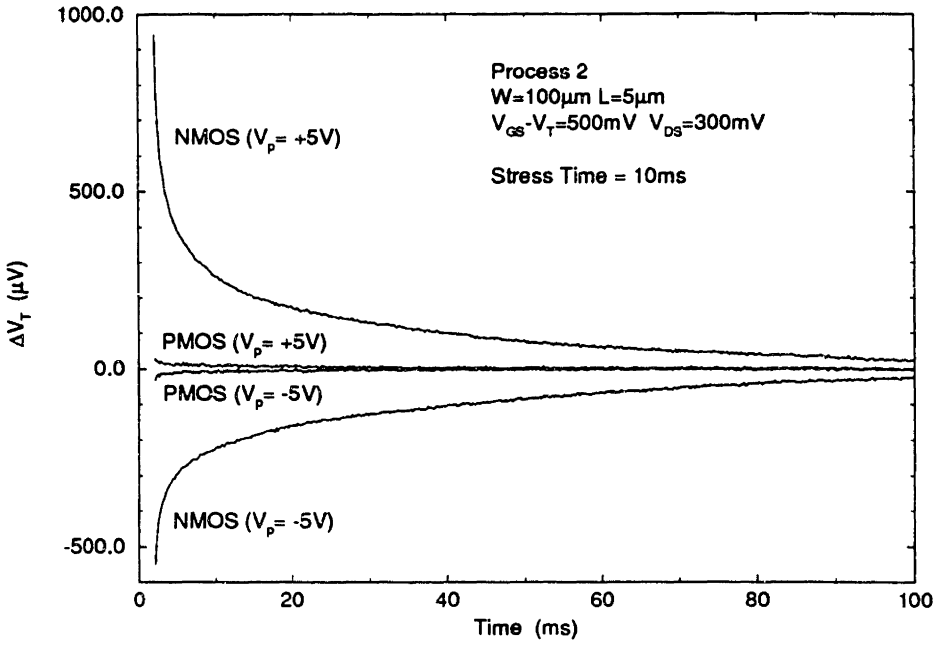
Figures 7.5 and 7.6 show linear region relaxation transients for NMOS and PMOS transistors in three different processes. The devices shown were stressed for 10ms with voltage pulses of amplitude ± 5V. Different processes show qualitatively similar behavior but exhibit quantitative differences with respect to amplitude of the V_T shifts, the relative magnitudes for positive and negative polarity stress voltages, and the relative magnitudes of the shifts in NMOS and PMOS devices. In NMOS devices, the amplitudes of the observed V_T shifts are generally found to be greater for positive than for negative stress voltages, but this depends on bias conditions as discussed in the following sections. In some cases the negative V_T shift may be nearly absent as in Fig. 7.6(a).

The response of PMOS devices shows a much greater variability among processes. Three general types of behavior are commonly observed as exemplified by Figs. 7.5(b) and 7.6(a),(b). In most processes the PMOS V_T shifts are about an order of magnitude less than those in NMOS. This type of behavior is exhibited by process 2 in Fig. 7.5. In this case the PMOS V_T shifts are nearly mirror images for positive and negative stress voltages, with peak values on the order of 20μV [Fig. 7.5(b)]. Larger values of PMOS V_T shifts are observed on some devices, as shown in Fig. 7.6. The shift may be larger for negative stress [Fig. 7.6(a)], but more frequently shows the opposite behavior [Fig. 7.6(b)].

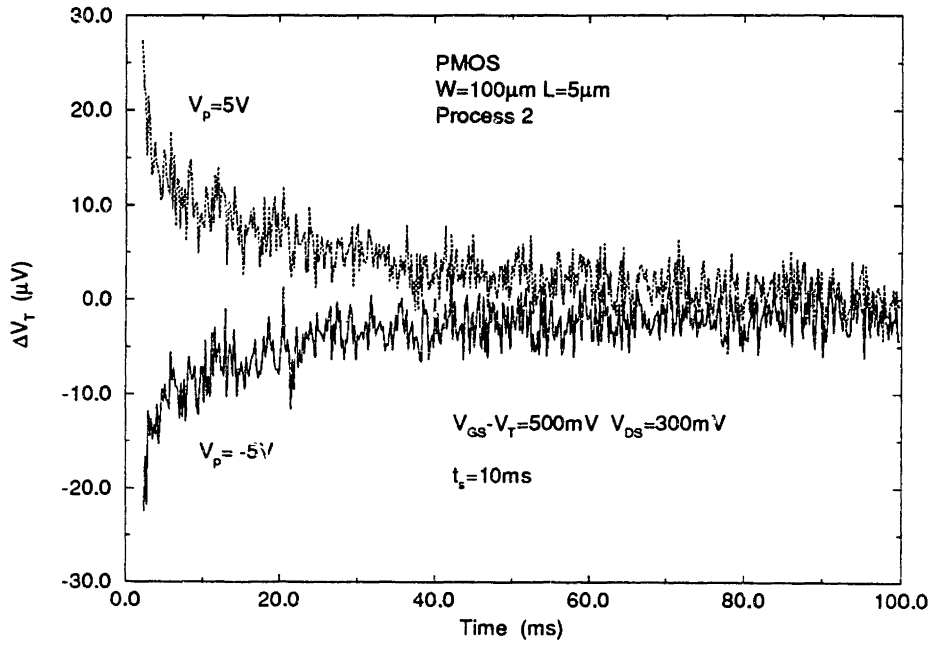
In order to determine the bias dependence of the V_T shifts we shall concentrate, in this section, primarily on NMOS devices due to their larger amplitude. The V_T shifts in PMOS devices are generally too small to be measured accurately at small stress voltages, and border on the resolution of the test circuit. An exception to this statement will be seen in the case of nitrided and reoxidized nitrided oxides, to be discussed in Sec. 7.2.3. Because of the large variations in threshold voltage shifts among processes and even among devices on the same wafer, we confine our attention to "typical" devices, as exemplified by Fig. 7.5(a).

When plotted on a logarithmic time scale the V_T relaxations are nearly linear for $t < t_s$ but shows some curvature for $t > t_s$ as shown in Fig. 7.8. The decay deviates markedly from the logarithmic decay curve in the range from from 1–100 ms, which includes most of the long-tail region of interest. This curvature is not predicted by previous theories but is consistent with Eqs. 4.70, 4.78 and 4.81 which predict deviations from a log decay approaching $1/t$ as shown in Fig. 4.11.

Figures 7.7(a) and (b) show relaxation transients for an NMOS transistor subjected to stress of duration 10μs and 10ms, respectively, at various voltage levels. The decay with time is nearly logarithmic. The voltage dependence is nearly linear for positive voltages but shows a marked nonlinearity for negative voltage stress, increasing very rapidly between -1V and -2V and then saturating for further increases in V_p in the negative direction.

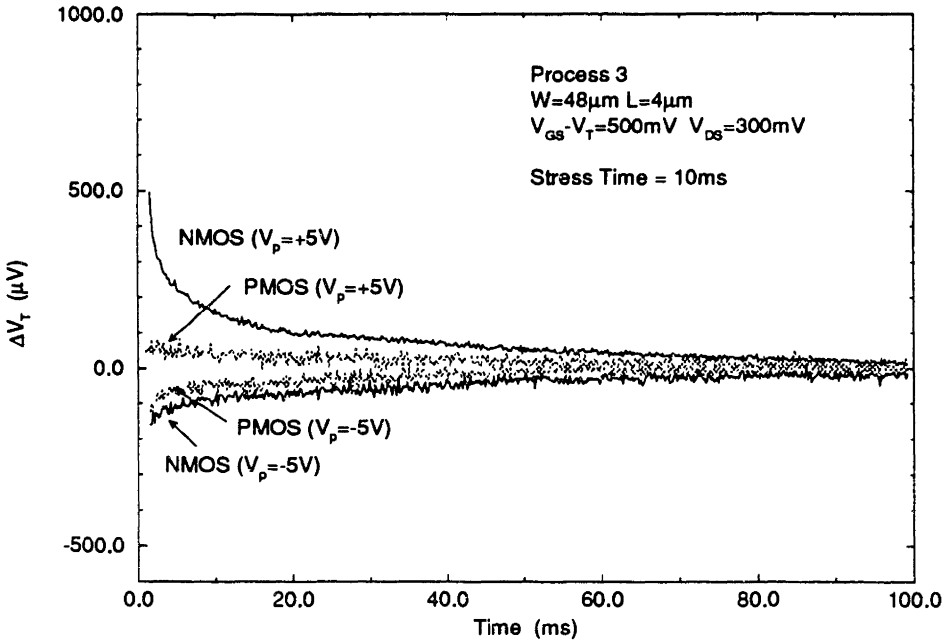


(a)

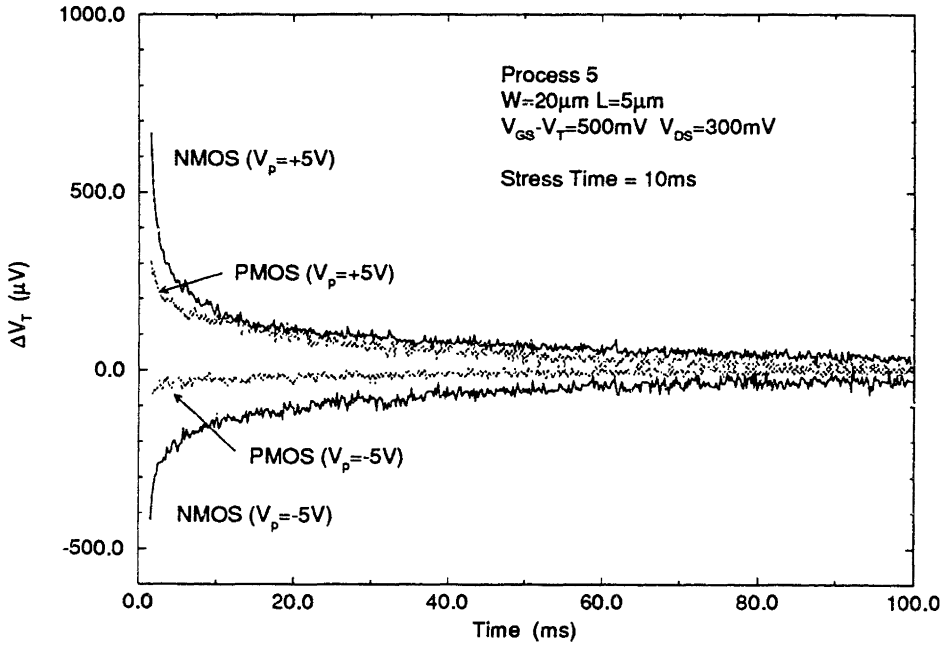


(b)

Figure 7.5: Relaxation transients for process 2 (a) comparison of NMOS and PMOS (b) close-up view of the PMOS V_T shift in (a).

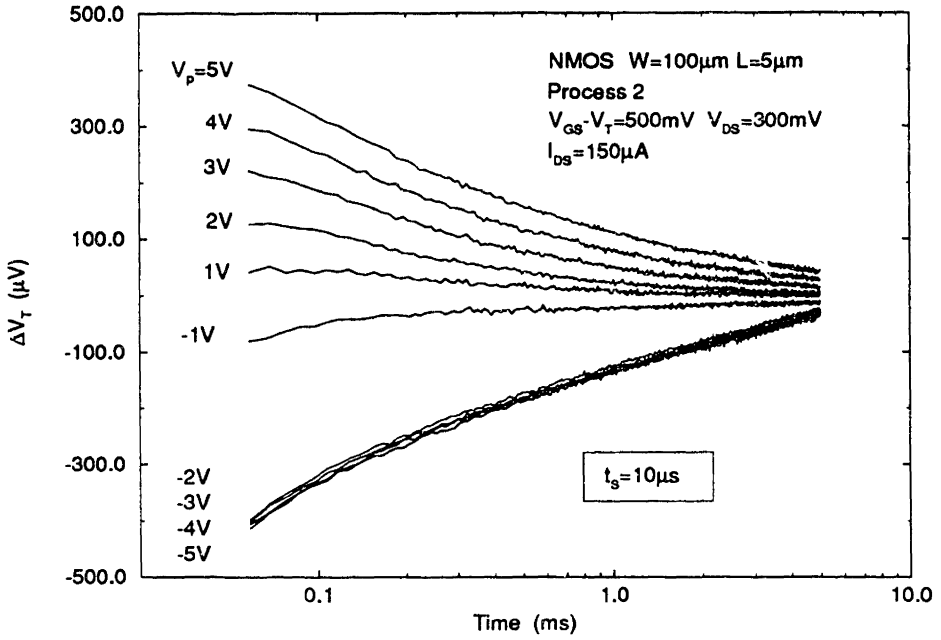


(a)

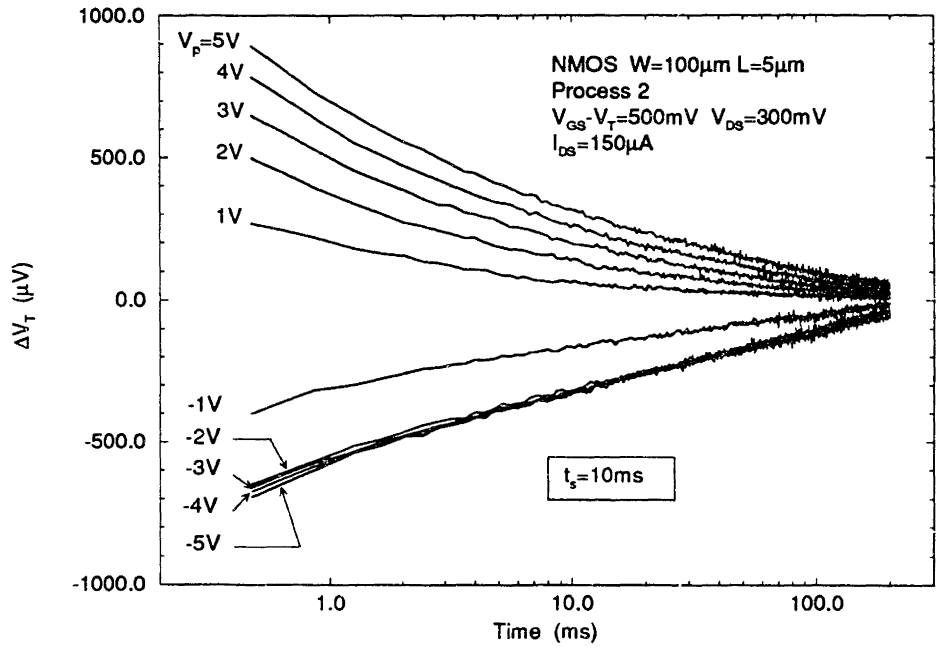


(b)

Figure 7.6: NMOS and PMOS relaxation transients for (a) process 3 (b) process 5.



(a)



(b)

Figure 7.7: Threshold voltage relaxation transients plotted on a log time scale with stress voltage as a parameter; (a) $t_s = 10\mu s$ (b) $t_s = 10ms$.

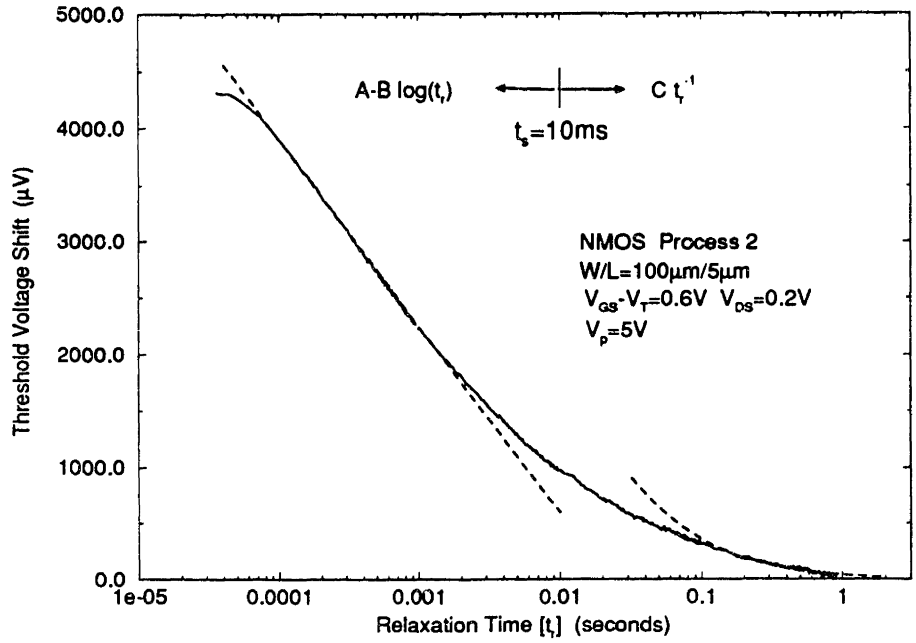


Figure 7.8: Relaxation measured with open loop circuit.

Stress Time

The threshold voltage shifts are observed to increase with stress time t_s as shown in Fig. 7.9. Note that even stress times on the order of $1\mu s$ can produce V_T shifts in excess of $100\mu V$ which persist for many milliseconds. When the initial threshold voltage shifts are plotted on a log time scale as in Fig. 7.10, the results generally fall on nearly straight line (process 4) but may also show some curvature (process 1). These results show good agreement with Eqs. 4.70, 4.78 and 4.81 which are plotted in Fig. 4.13. The curvature, which is not predicted by conventional tunneling models, results from the it/cb component [Eq. 4.78] which varies as $\ln^2(t_s)$. Note that, in the measurement of the stress time dependence, t_r is fixed and satisfies $t_r \ll t_s$ in Eq. 4.78. A large amount of curvature could be explained by a large interface state density which enhances the it/cb component. It can also be explained, of course, by a larger lr/cb component due to lattice relaxation.

Figure 7.11 plots the initial threshold voltage shift as a function of stress time for NMOS transistors in processes 2 and 5, with stress voltage as a parameter. For positive stress the V_T shift increases nearly logarithmically with stress time and increases proportional to stress voltage as expected from the theory [Figs. 4.13, 4.14]. Also evident is the sharp nonlinearity for negative stress voltages followed by saturation of the V_T shift. The stress time dependence is also weaker for negative voltages. This nonlinearity is not predicted by the approximate strong inversion model of Sec. 4.4

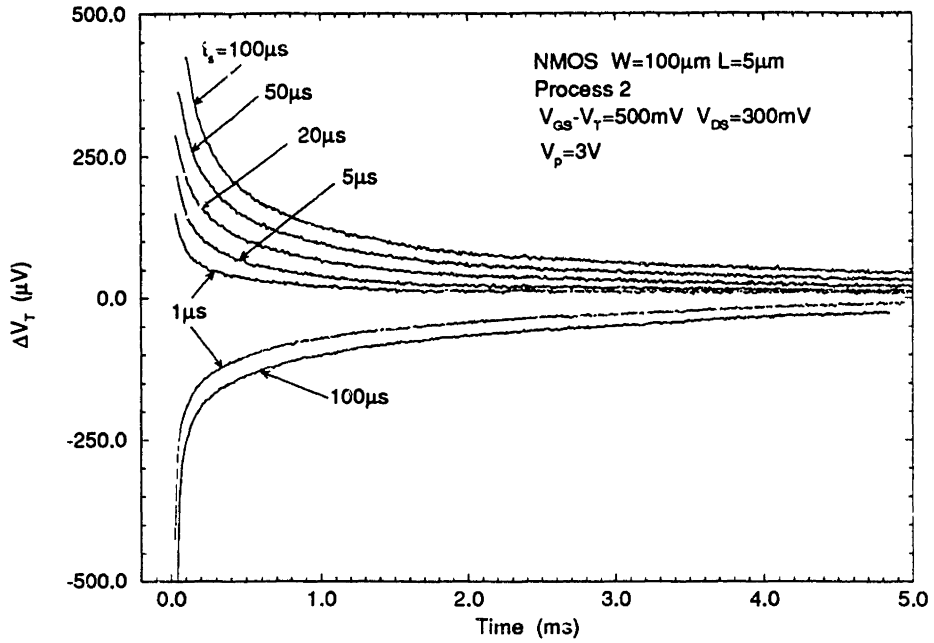


Figure 7.9: Dependence of threshold voltage relaxation on stress time.

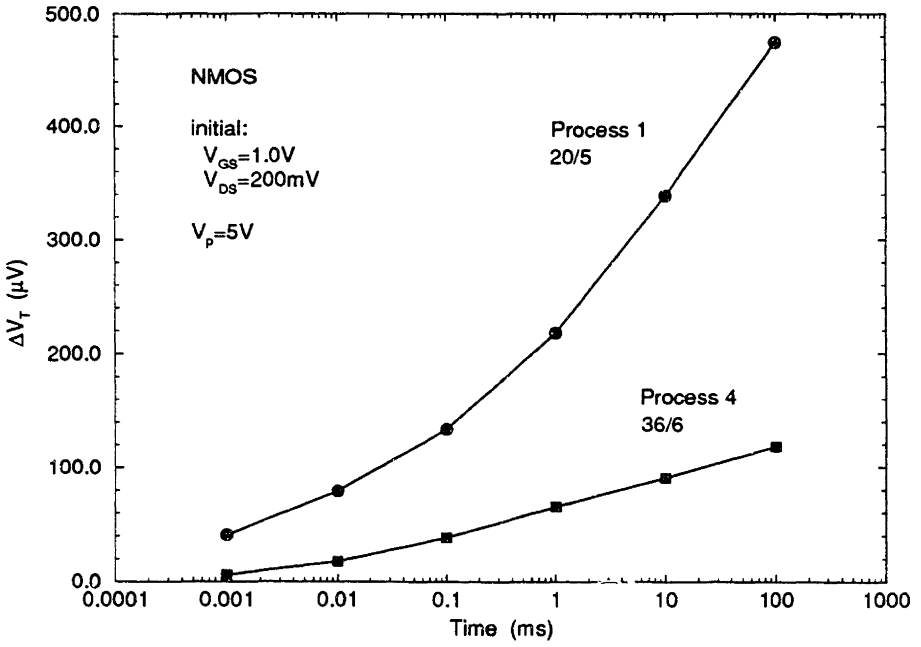


Figure 7.10: Initial threshold voltage shift vs. stress time for two different processses.

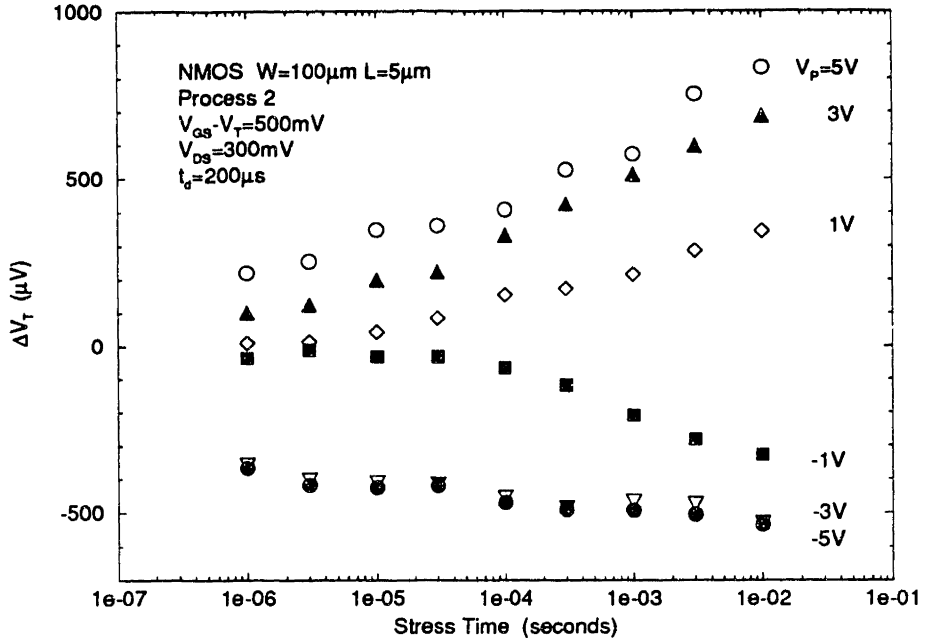


Figure 7.11: Initial amplitude of threshold voltage shift versus stress time.

and requires the full numerical model of Sec. 7.2.4.

Pulse Amplitude

In order to investigate the dependence of the V_T shifts on pulse amplitude, we plot in Fig. 7.12 the initial amplitude of the V_T shift versus stress voltage. At a stress time of 10ms, the V_T shift increases nearly linearly with stress voltage in the positive direction as noted previously. The magnitude of the V_T shift for negative voltages initially increases proportional to V_p, but tends to saturate for $V_p < -2$V. At shorter stress times, e.g. $t_s = 10\mu$s, the V_T shift increases less rapidly with voltage in the positive direction and makes an abrupt transition to a saturation level of slightly smaller amplitude. All of the processes characterized showed a similar stress voltage dependence.

Figure 7.13(a) shows the ΔV_T–V_p characteristic for an NMOS transistor in process 2 at a stress time $t_s = 10\mu$s. Also shown in Fig. 7.13(b) is the calculated surface potential, normalized to kT/q, versus stress voltage. With $V_T = 0.7$V and an initial gate-source voltage of $V_{GS} = 1.2$V, the device is initially in strong inversion, near the upper knee of the surface potential curve when $V_p = 0$. Thus, in the region where ΔV_T increases with V_p, the surface is strongly inverted with E_F pinned near the conduction band. For small negative voltages, $-0.7 < V_p < 0$, the device remains in inversion with $U_s > 2U_F$ and ΔV_T decreases proportional to V_p. For $-1.4V < V_p < -0.7V$ the

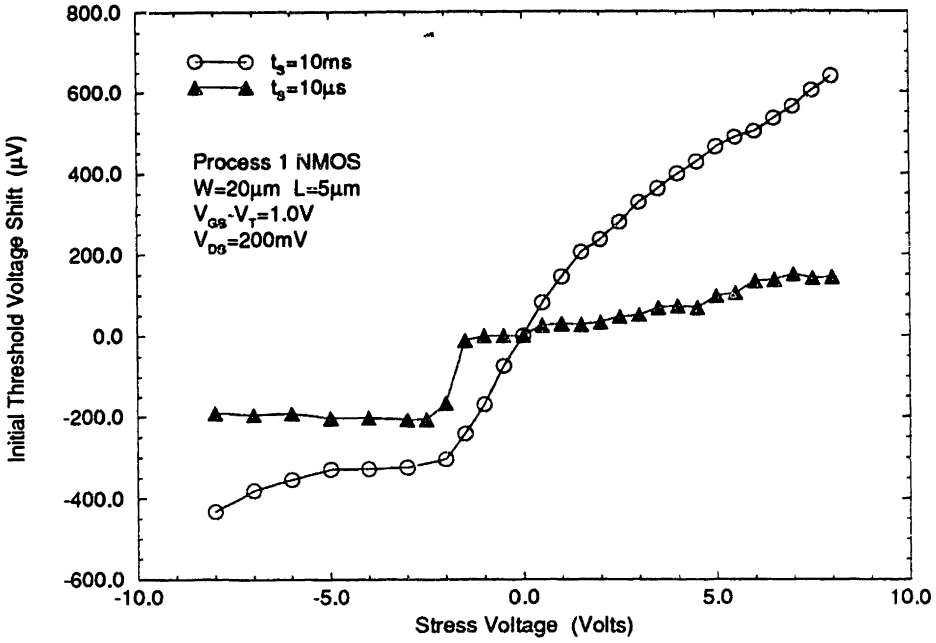


Figure 7.12: Initial V_T shift versus stress voltage for Process 1.

surface is weakly inverted and the surface potential decreases rapidly with increasingly negative V_p. In this region the Fermi level sweeps abruptly through the bandgap. For $V_p < -1.4V$ the surface accumulates and at $V_p \approx -2V$ the band bending becomes pinned with the Fermi level near the top of the valence band. This region, $V_p < -2V$ corresponds to the region in which ΔV_T saturates with decreasing V_p.

Figure 7.13(c) shows the change in power dissipation which occurs during stress, as a function of pulse amplitude V_p. For these particular bias conditions the power dissipation actually decreases during positive stress, thereby tending to cool the DUT. The fact that the V_T shift is positive provides further evidence that it cannot be due to thermal effects. It should be noted from this figure that the point where the current cuts off, at the minimum of the ΔP_d curve, occurs at $V_p \approx -0.7$V, when the device comes out of strong inversion, whereas the threshold voltage shift continues to decrease until the knee of the surface potential is reached at $V_p \approx -2V$. This implies that the saturation phenomenon is not current-related but depends on the position of the electron quasi-Fermi level in the bandgap.

The above considerations lead to the following interpretation in terms of the tunneling model. In strong inversion, i.e. positive and slightly negative stress voltages $V_p > V_p(2U_F) \approx -0.7$V, the surface is strongly inverted and the Fermi level is pinned near the conduction band. Thus, the it/it component is disabled and the V_T shift is due almost entirely to the cb/cb and it/cb components which increase with voltage according to Eqs. 4.78 and 4.81. For $-2V < V_p < -0.7V$, the Fermi level sweeps

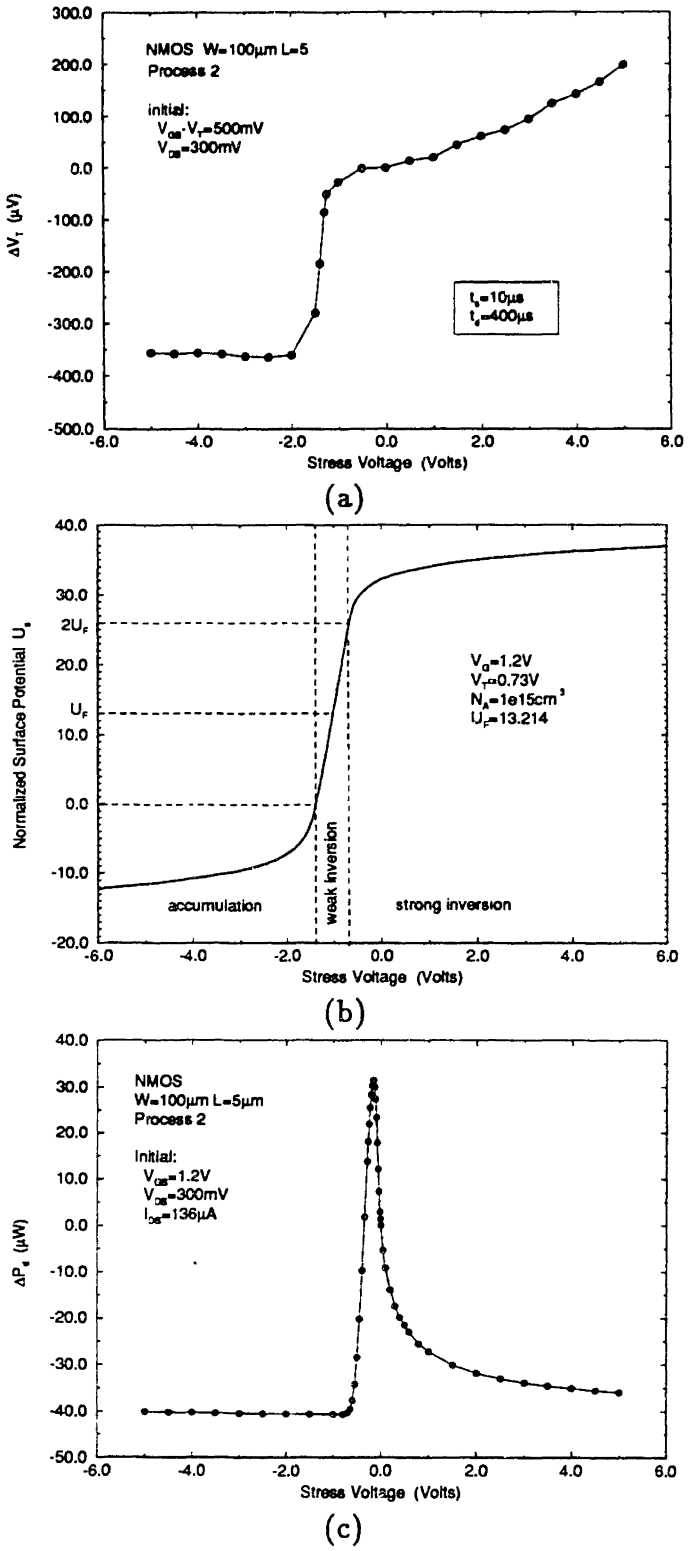


Figure 7.13: (a) Stress voltage dependence (b) Surface potential versus gate voltage (c) Change in power dissipation during stress [Process 2].

through the bandgap and two-step tunneling between oxide traps and fast surface states dominates. this mechanism is suppressed when the Fermi level becomes pinned near the valence band for $V_p < -2V$.

For increasing negative voltages, the pinning of the Fermi level near the valence band explains why there is no further contribution due to traps within the bandgap, but does not explain why there is no response due to traps below the valence band. These traps are initially filled with electrons and would be raised above the Fermi level during stress. This should lead to *vb/vb* and *it/vb* components analogous to the *cb/cb* and *it/cb* components observed for positive pulses. Clearly, this does not occur, as shown in Fig. 7.13(a), which indicates that the threshold voltage shift clamps in the negative direction. This behavior is observed using both the open-loop [Fig. 6.2] and closed-loop [Fig. 6.3] measurement techniques.

Two explanations suggest themselves: Either (1) no hole trapping occurs during negative stress or (2) hole trapping occurs, but the rate at which trapped charge is reemitted when the surface returns to inversion occurs too rapidly to be observed. Since the surface accumulates during large negative pulses, holes will be captured if traps are present below the valence band. Thus the first possibility leads us to the conclusion that there are no traps below the valence band. The difficulty with this explanation is that PMOS transistors exhibit exactly the opposite behavior as NMOS, saturating for positive voltages [1]. The saturation of the threshold voltage shift for accumulating pulses on both types of substrate would imply that NMOS devices have no traps below the valence band, while PMOS devices have no traps above the conduction band. There is no justification for believing that the trap distributions can be so radically different in oxides grown on substrates which differ only in dopant type.

The reason for the saturating behavior must therefore lie in the second explanation. The question immediately arises as to why the emission of trapped charge should occur so rapidly after an accumulating pulse (negative in NMOS, positive in PMOS) and so slowly after an inverting pulse (positive in NMOS, negative in PMOS). Consider the energy band diagrams in Fig 7.14 which show the measurement sequence for an NMOS transistor during positive (top) and negative (bottom) gate voltage stress. We shall assume a model in which traps above the conduction band are acceptor-like (neutral when empty, positively charged when occupied by an electron) and below the valence band are donor-like (neutral when filled by an electron and positively charged when empty). There is ample research to support a model of this type [157]. We neglect traps within the bandgap for now, which are not relevant to the present discussion. The transistor is initially in inversion, and we assume that all traps above the Fermi level are empty (neutral) and below the Fermi level are filled (neutral). During a positive (inverting) gate voltage pulse, traps above the conduction band are

[1] While this effect is too small to be measured in conventional SiO_2 dielectrics, it can be readily observed in nitrided oxides as discussed in Sec. 7.2.3.

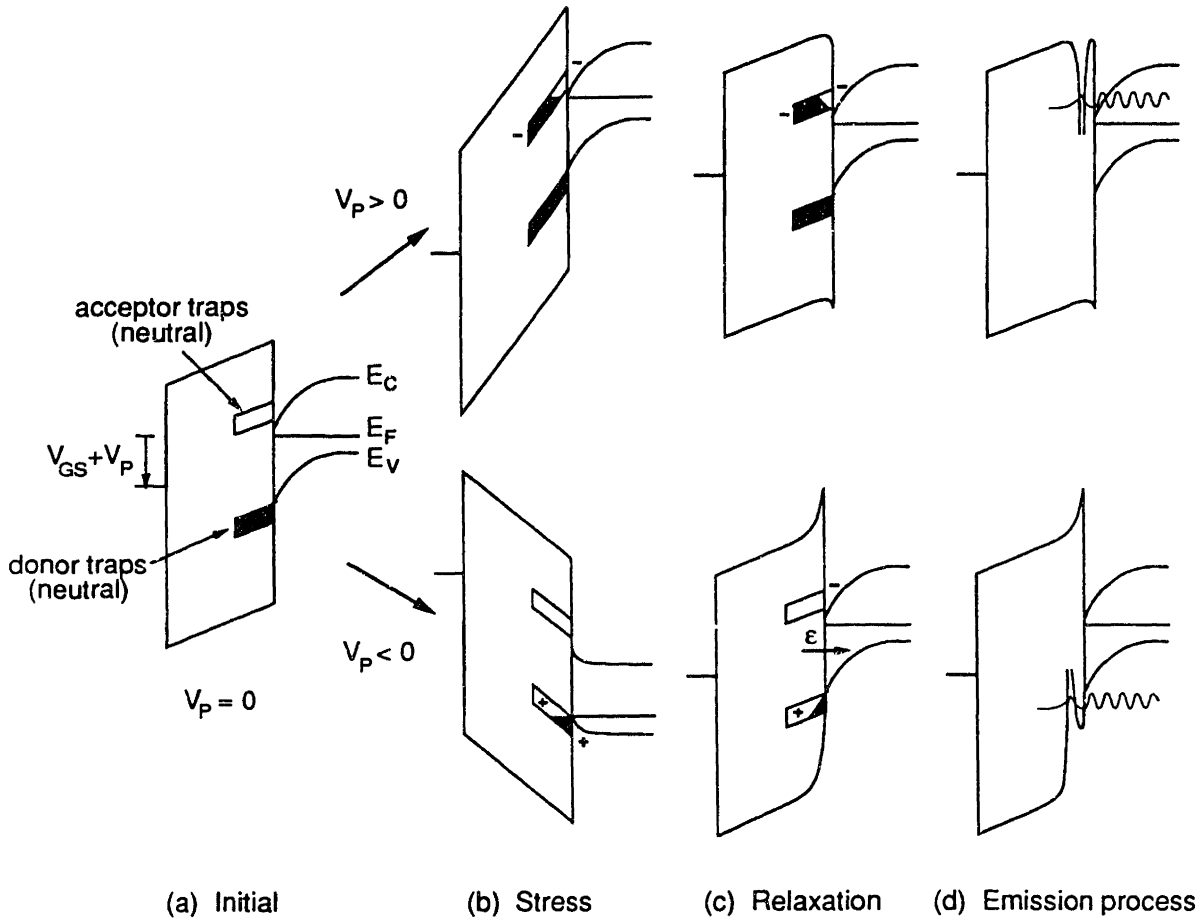


Figure 7.14: Energy band diagrams for positive (top) and negative (bottom) stress sequences on an NMOS transistor, showing field emission of holes trapped during negative stress.

lowered below the Fermi level where they become filled with electrons as shown in Fig. 7.14(b). When the gate voltage returns to its initial value the electric field due to the negatively charged traps will cause some rounding of the barrier as shown in (c). Since the charge on both sides of the barrier is negative at all times during the measurement, it will be imaged on the gate, and the barrier rounding will be small. The trapped electron tunnels back out through essentially a trapezoidal barrier as discussed in Chapter 3.

Now consider the case in which a negative (accumulating) pulse is applied to the gate. The Fermi level becomes pinned in the valence band and the surface is flooded with holes. Traps below the valence band are shifted above the Fermi level by the electric field in the oxide, where they may capture holes (emit electrons) becoming positively charged [Fig. 7.14(b)]. When the transistor returns to inversion

[Fig. 7.14(c)], the surface is again flooded with electrons. Intuitively, a recombination of the electrons and holes appears possible. However, since the inversion layer electrons and the trapped holes are separated by an insulating barrier, they can only recombine by tunneling, and this is impeded because they are separated in energy by the silicon bandgap. It is possible however, for trapped holes to tunnel back out to filled valence band states in the semiconductor. The difference between this case [Fig. 7.14(c) bottom] and the previous case [Fig. 7.14(c) top] is that there are now trapped holes on one side of the interface and inversion layer electrons on the other. The dipole field which originates on a trapped hole and terminates on an inversion layer electron, can result in a high local electric field in the vicinity of the trapped hole, as shown in Fig. 7.14(d). Under this high electric field, detrapping can occur by Frenkel-Poole emission or field (tunneling) emission.

Furthermore, since the transistor returns to inversion with the Fermi level high in the bandgap, there is a very high density of filled electron states (empty hole states) below the valence band into which the trapped holes can tunnel. In the inverting case [Fig. 7.14(c) top], the Fermi level is always near the conduction band and the density of empty states into which the trapped charge can tunnel is lower.

This explanation for the saturating behavior of the threshold voltage shift holds equally well for positive pulses in PMOS devices. In this case, the charge on both sides of the barrier is always positive during negative gate voltage pulses, but forms a dipole layer across the interface after stress with a positive pulse. Reemission occurs under the high local fields originating on inversion layer holes and terminating on trapped electrons.

Ideally, the closed-loop measurement circuit of Fig. 6.3, forces a constant current and hence a constant electric field in the silicon throughout the measurement sequence. In the steady state the average oxide field near the interface will therefore also be constant and any changes in trapped charge will simply change the amount of charge imaged on the gate, thereby shifting the threshold voltage. In practice, however, the feedback loop requires a finite time to respond after the stress period, and during this time the fields near the interface may become high. Furthermore, while the *average electric field* may not be excessive, the *local electric field* in the immediate vicinity of a trapped charge can become very high. For an electron trapped 1-10Å from the interface, the local field is on the order of 10^6–10^8 V/cm.

Since high-field effects were not taken into account in the quantitative model, our simulation results are not valid in accumulation. This implies, for example, that threshold voltage relaxation measurements cannot be used to extract trap densities below the valence band in NMOS or above the conduction band in PMOS devices. Since most of our measurements are carried out for inverting pulses we shall not consider the field-emission mechanism in any further detail, although it is a worthwhile topic for further research.

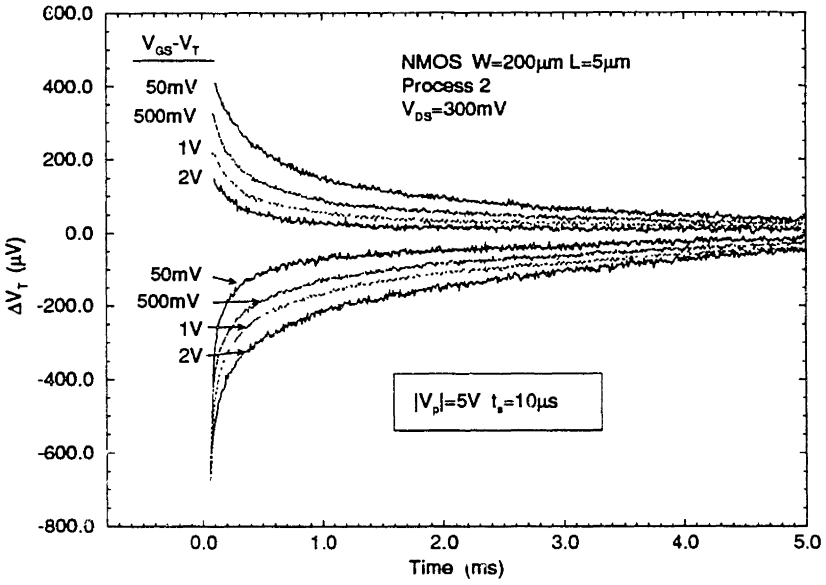


Figure 7.15: Threshold voltage relaxation transients with initial gate-source voltage as a parameter.

Initial Gate Bias

As the pre-stress gate bias level is varied, the V_T relaxation transients shift as shown in Fig. 7.15, plotted for an NMOS transistor subjected to $10\mu s$ of stress of amplitude ± 5V. As the initial level of inversion $V_{GS} - V_T$ increases, the threshold voltage relaxation curves shift toward more negative voltages. That is, the amplitude of the shift for positive stress decreases and the amplitude of the shift for negative stress increases as the initial gate bias is increased.

This is shown more clearly in Fig. 7.16 which plots the initial amplitude of the V_T shift versus stress voltage parameterized by the pre-stress value of $V_{GS} - V_T$. Similar behavior is observed in all the processes studied. Figure 7.17 shows the analogous plots for an NMOS device in process 1 at stress times of $10\mu s$ and 10ms.

The explanation of this behavior in the tunneling model is straightforward. In the *it/it* mechanism, traps above the Fermi level contribute to threshold voltage shifts under positive stress and traps below E_F contribute for negative stress. As the initial level of inversion increases, the Fermi level starts out higher in the bandgap initially and sweeps over a wider range of trap energies before becoming pinned at the valence band, giving an increased threshold voltage shift for negative gate voltage pulses. The V_T shift for positive voltage pulses decreases correspondingly.

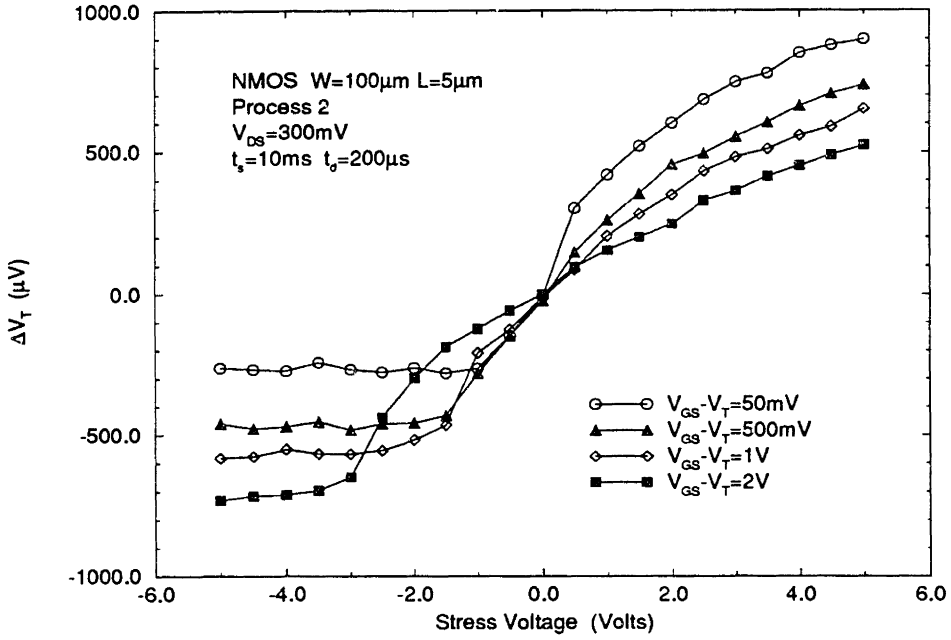


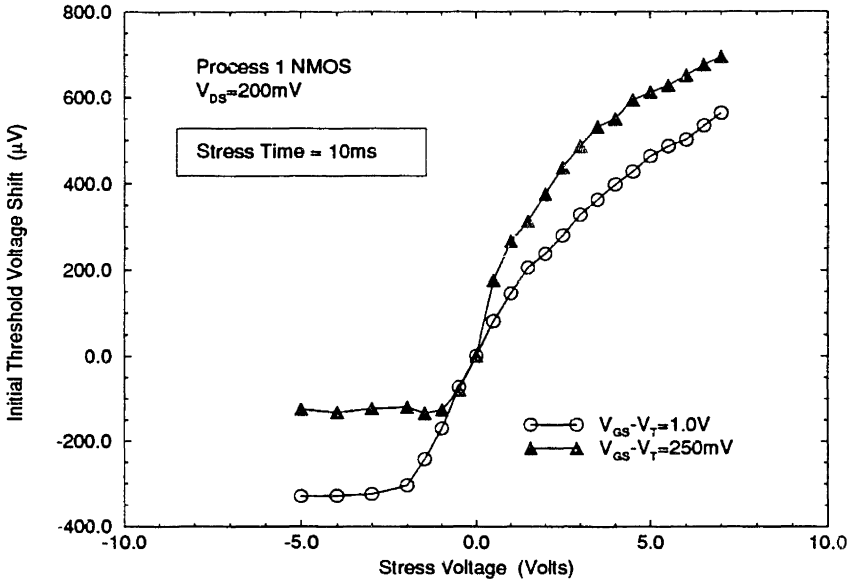
Figure 7.16: Initial threshold voltage shift vs. stress voltage with initial gate-source voltage as a parameter.

Substrate Bias

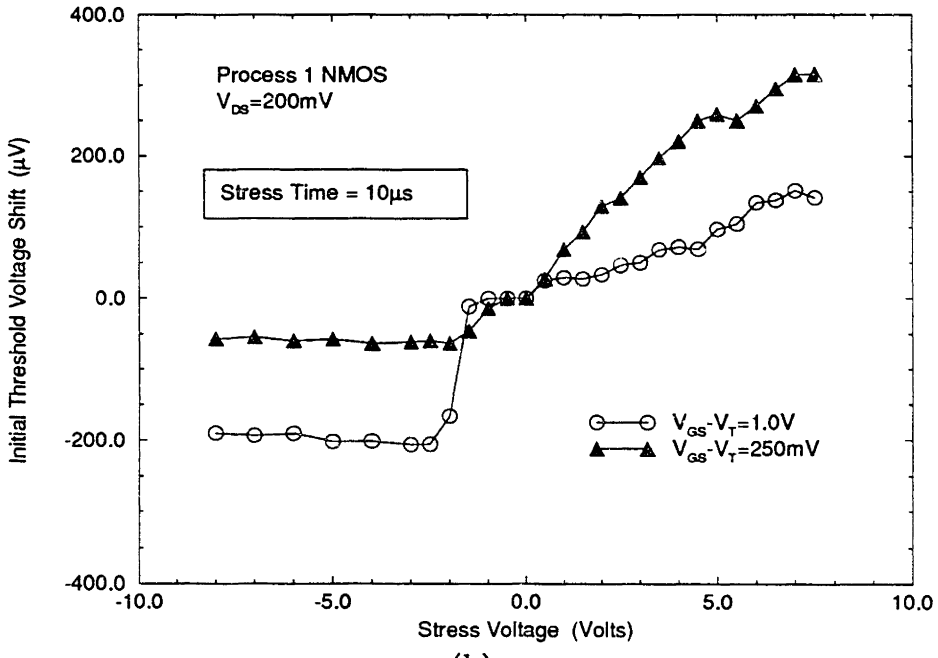
Fig. 7.18 shows NMOS V_T relaxation transients for $V_{BS} = 0$ and $V_{BS} = -5$V. The presence of a backgate bias nearly eliminates the V_T shift for -5V stress while the shift for +5V stress is very slightly reduced. In these experiments, the initial gate voltage level is adjusted to maintain a constant current level as the source-substrate bias is adjusted.

The features of the backgate bias dependence are brought out more clearly in Fig. 7.19 which plots the initial threshold voltage shift versus pulse amplitude for several values of V_{BS}. It is apparent that the shift for negative gate voltage stress is still present for non-zero backgate biases but the position of the step simply shifts to more negative voltages by an amount proportional to V_{BS}. The slight decrease in ΔV_T for positive pulses is also evident in this figure.

When a MOSFET is biased with a non-zero backgate bias, the electron and hole quasi-Fermi levels separate as shown in Fig. 7.20. Figure 7.20(a) shows the initial band diagram, with V_{GS} adjusted to strongly invert the surface. At this point, the electron quasi-Fermi level E_{Fn} is near the conduction band and all fast surface states below E_{Fn} are essentially filled. A negative gate voltage pulse raises the energies of all surface states with respect to the electron quasi-Fermi level as shown in Fig. 7.20(b). Given sufficient time, traps above E_{Fn} will be mostly empty. However, the rate at



(a)



(b)

Figure 7.17: Initial threshold voltage shift vs. stress voltage with initial gate bias as a parameter [Process 1] (a) $t_s = 10$ms (b) $t_s = 10\mu$s.

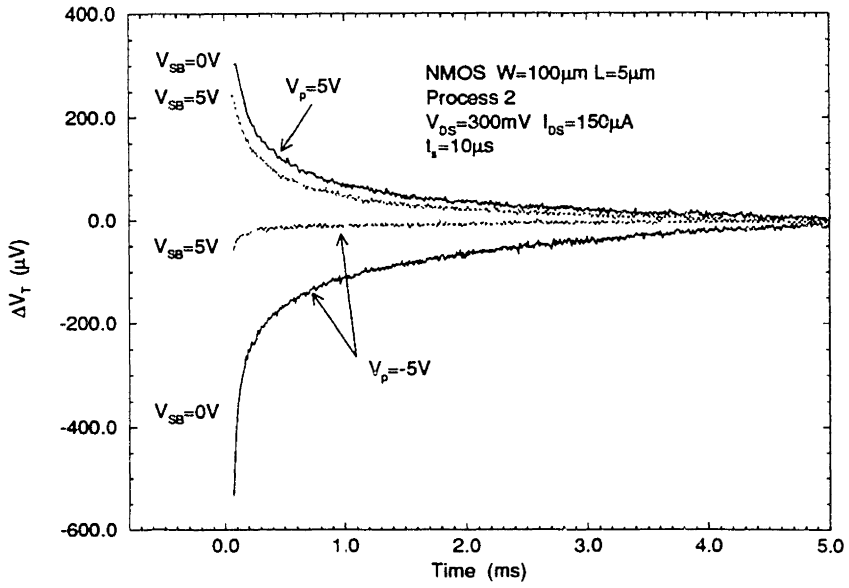


Figure 7.18: Threshold voltage relaxation transients at $V_{BS} = 0$V and $V_{BS} = -5$V.

which this occurs is determined by the interface state time constants which depend in turn on the surface potential. Detrapping can occur by electron emission or by hole capture. The time constant for electron emission increases exponentially with temperature and decreases exponentially with energetic depth below the conduction band according to Eqs. 3.53 and 3.71. Thus electron detrapping from traps near and below midgap must occur by hole capture. However, during negative gate voltage pulses, the band bending must pass through a depletion region in which no holes are present. The range of gate voltages over which the surface is depleted increases in proportion to the backgate bias. When sufficient negative bias is applied to the gate, as in Fig. 7.20(c) to accumulate the surface, electrons are rapidly swept out of surface states by recombination with the abundance of holes. An explanation in terms of the lattice relaxation model is analogous since both time constants have the same form in the thermally-limited regime [Sec. 4.9].

It follows that, if the abrupt steps in the ΔV_T–V_p characteristics are due to the passage of the Fermi level through a depletion region followed by the sudden detrapping of electrons in accumulation, that the features of this step should show some variation with time and temperature.

Indeed, we have seen in Figs. 7.12 and 7.17 and shall see again in Figs. 7.33 and 7.36, that increasing the stress time makes the step more gradual, extending the region of proportionality to V_p all the way down to the voltage at which the Fermi level becomes pinned at the valence band edge. As the stress time increases relative to

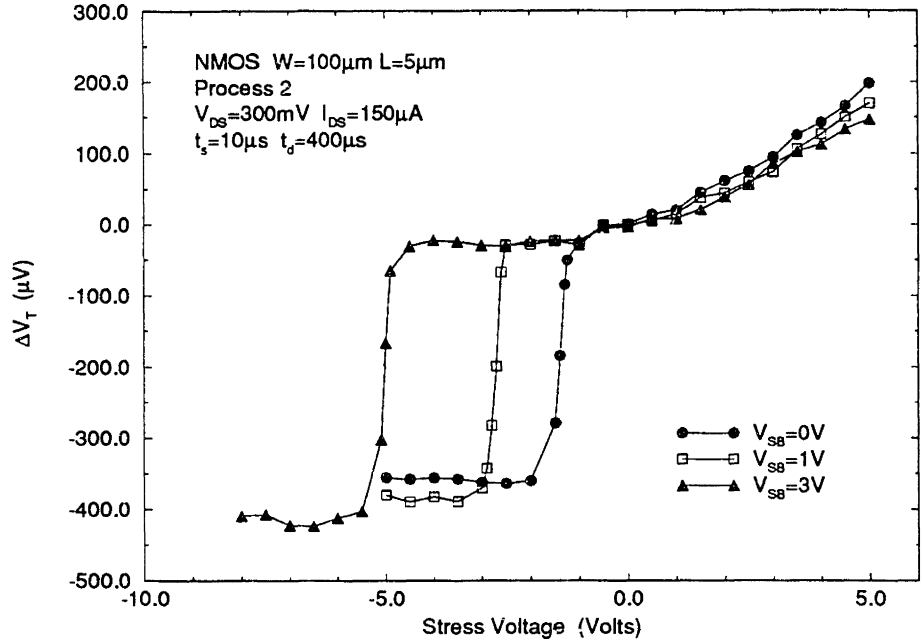


Figure 7.19: Initial threshold voltage shift vs stress voltage with substrate bias as a parameter.

the time constants for electron emission [Fig. 7.17(a)], the trapped charge can follow the applied gate voltage, even for negative pulses. When the pulse duration is too short [Fig. 7.17(b)] interface traps retain their charge until the pulse is sufficient to drive the surface into accumulation, at which point the trapped electrons recombine all at once with accumulation layer holes. Increasing the temperature also tends to broaden the abrupt step by reducing trap time constants, as discussed below.

The pulse voltage V_p required to accumulate the surface of an NMOS transistor (p-substrate) can be easily calculated. Kirchoff's voltage law requires

$$V_{GB} - V_{FB} = \psi_s + \frac{Q_b}{C_{ox}}$$

where Q_b is the magnitude of the depletion charge. Since $V_{GB} = V_{GS} - V_{BS}$,

$$V_{GS} = \psi_s + V_{FB} - V_{SB} + \frac{Q_b}{C_{ox}} \tag{7.1}$$

Suppose that when $V_{SB} = 0$, accumulation occurs when the surface potential reaches $\psi_s = \phi_{acc} < 0$. The corresponding gate voltage is

$$V_{GS}|_{acc} = \phi_{acc} + V_{FB} + \frac{Q_b(\phi_{acc})}{C_{ox}}$$

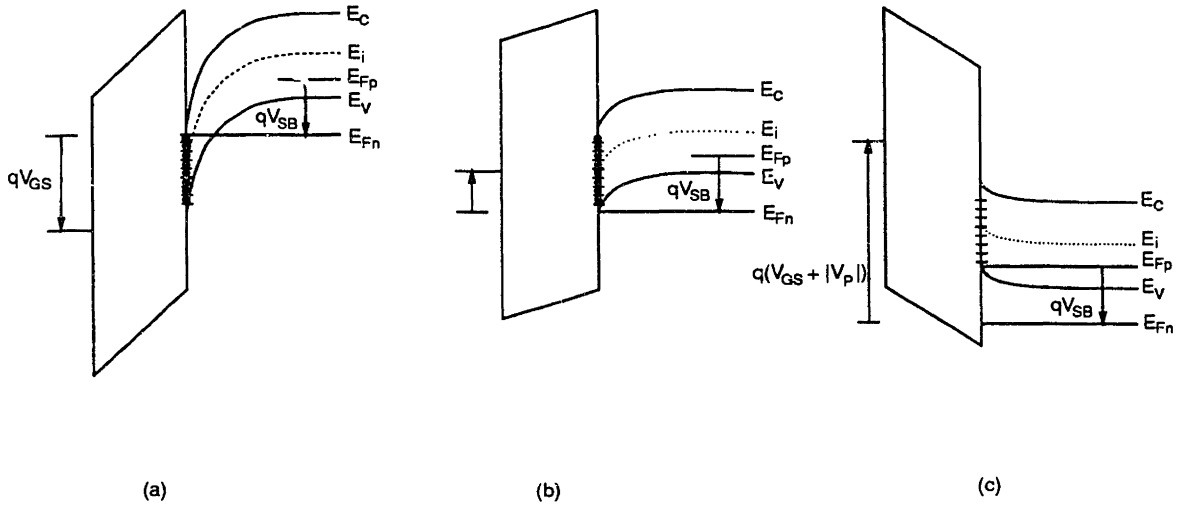


Figure 7.20: Band diagram for an NMOS transistor with $V_{SB} > 0$ (a) pre-stress $V_p = 0$ (inversion) (b) small negative stress V_p (depletion) (c) large negative stress V_p (accumulation).

We assume that the device is initially in inversion with

$$V_{GSi} = V_{T0} = V_{FB} + 2\phi_F + \frac{Q_b(2\phi_F)}{C_{ox}}$$

The change in gate-source voltage $V_p = V_{GS}|_{acc} - V_{GSi}$ that causes the surface to accumulate is therefore

$$V_p|_{V_{SB}=0} = \phi_{acc} - 2\phi_F + \frac{Q_b(\phi_{acc})}{C_{ox}} - \frac{Q_b(2\phi_F)}{C_{ox}}$$

When the substrate is tied high, $V_{SB} > 0$, the initial gate-source voltage is adjusted to give the same initial current as when $V_{SB} = 0$. Since the electron quasi-Fermi level at the source is now displaced from the bulk quasi-Fermi level by $-V_{SB}$, the same initial level of inversion now requires $\psi_s = 2\phi_F + V_{SB}$. The initial gate-source voltage is therefore

$$V_{GSi} = V_T = V_{FB} + 2\phi_F + \frac{Q_b(2\phi_F + V_{SB})}{C_{ox}}$$

Accumulation still occurs at $\psi_s = \phi_{acc}$ since the hole quasi-Fermi has not shifted with respect to the bulk. From Eq. 7.1,

$$V_{GS}|_{acc} = \phi_{acc} + V_{FB} - V_{SB} + \frac{Q_b(\phi_{acc})}{C_{ox}}$$

The pulse voltage required to achieve accumulation is now

$$V_p|_{V_{SB}>0} = \phi_{acc} - 2\phi_F - V_{SB} + \frac{Q_b(\phi_{acc})}{C_{ox}} - \frac{Q_b(2\phi_F + V_{SB})}{C_{ox}}$$

Thus, a substrate bias causes the point of accumulation to shift along the V_p axis by an amount

$$\begin{aligned}\Delta V_p &= V_p|_{V_{SB}>0} - V_p|_{V_{SB}=0} \\ &= -V_{SB} - \frac{1}{C_{ox}}[Q_b(2\phi_F + V_{SB}) - Q_b(2\phi_F)] \\ &= -V_{SB} - \gamma\left[\sqrt{2\phi_F + V_{SB}} - \sqrt{2\phi_F}\right] \end{aligned} \tag{7.2}$$

where the body-effect parameter is defined by

$$\gamma = \frac{1}{C_{ox}}\sqrt{2q\epsilon_{si}N_A} \tag{7.3}$$

The second term in Eq. 7.2 is just the increase in threshold voltage due to the body-effect with a substrate voltage V_{SB}. This shift of the characteristics along the V_p axis, which is equal to the substrate bias plus the increase in threshold voltage, can be seen in Fig. 7.19.

Temperature

Measurements of the temperature dependence of the V_T shifts are made with the circuit of Fig. 6.3 by simultaneously heating or refrigerating both the DUT (M1) and the reference device (M2). The remainder of the circuitry is maintained at room temperature. The pre-stress value of the drain current and hence, drain voltage, vary with temperature. By adjustment of the circuit variables, the measurements can therefore be carried out either under conditions of constant current or constant drain voltage, but not both.

Figure 7.21 shows V_T relaxation transients for an NMOS transistor stressed at $\pm$5V for 10ms at several different temperatures at constant V_{DS}. The amplitude of the V_T shifts for both positive and negative stress voltages decrease slightly with increasing temperature. Figures 7.22(a) and (b) plot the initial amplitude ($t_d = 0.57\mu$s) of the threshold voltage shift versus temperature under conditions of constant current and constant drain voltage, respectively. In either case, the amplitude of the V_T shift decreases with increasing temperature. Similar behavior was found in all the devices tested as shown in Fig. 7.23. For positive stress the temperature coefficient at constant V_{DS} is nearly linear and varies from -0.1μV/°C to -1.25μV/°C.

It was shown in Sec. 4.5 that, for positive gate voltage pulses in an NMOS transistor, a decreasing threshold voltage shift with increasing temperature can result from the movement of the Fermi "probe" Δf to lower energies, where the trap density may be less. It can also be explained by the correlated mobility term in Eq. 4.104 which varies as $T^{-3/2}$ [Sec. 4.2]. For negative pulses, the decrease in amplitude arises through mobility fluctuations as well as through the activated nature of the trapping

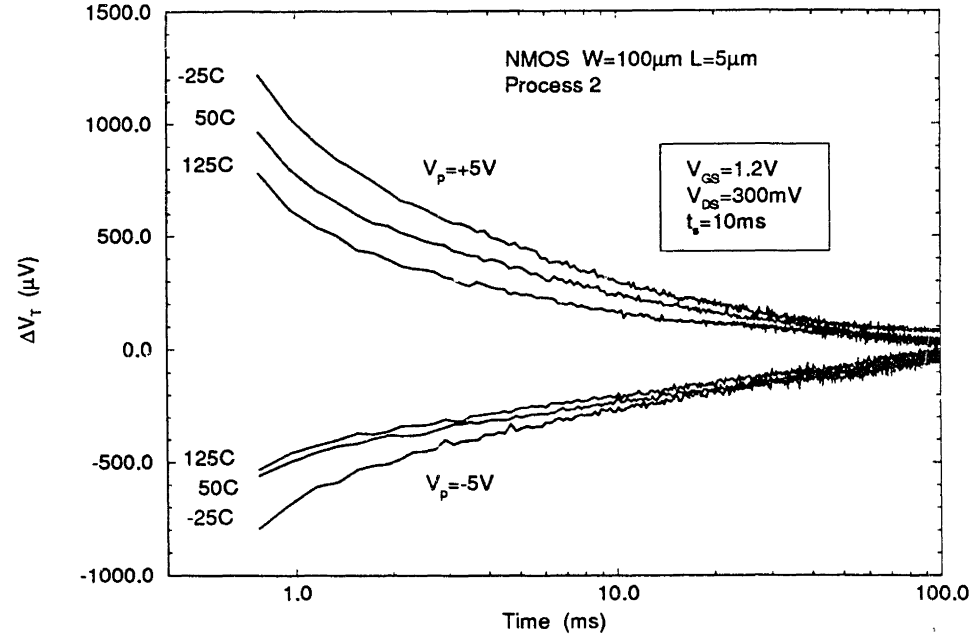
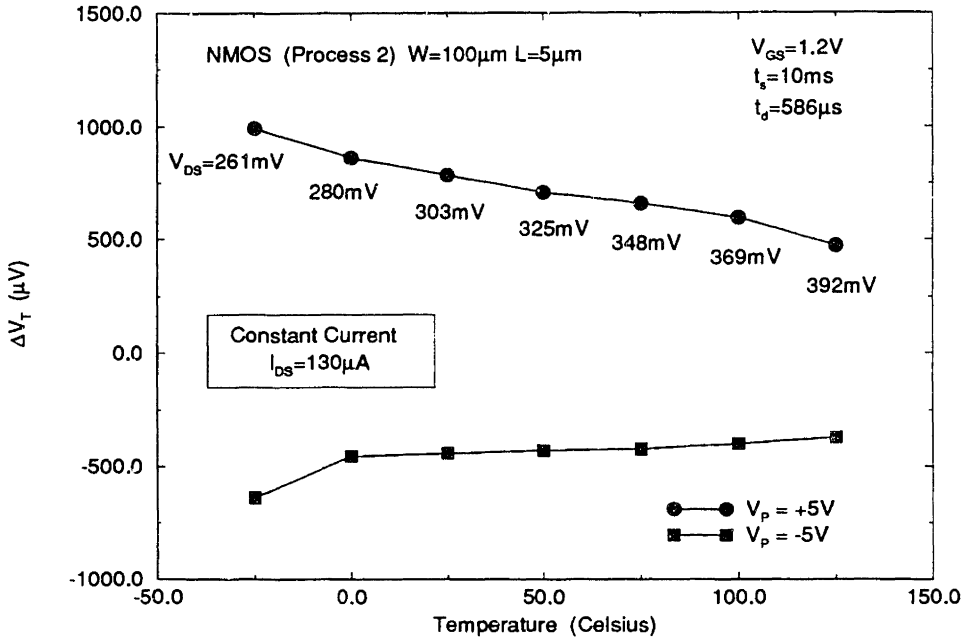


Figure 7.21: Threshold voltage relaxation transients at several different temperatures.

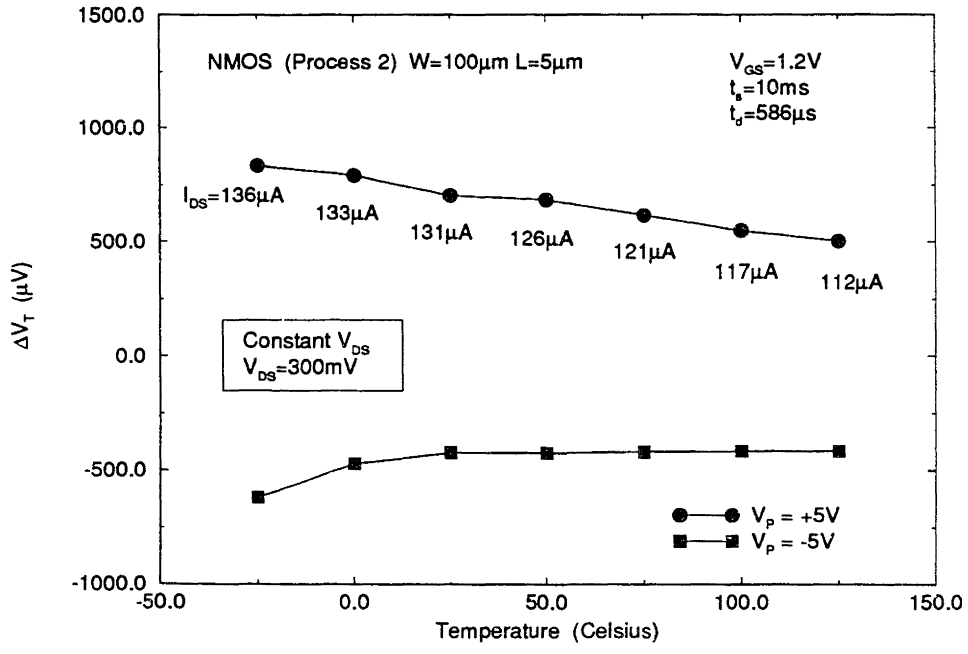
mechanism, whether lattice relaxation or two-step. This is discussed further below, for the case of the two-step mechanism. Lattice relaxation follows by analogy since the thermal part of the time constant has the same form in either case [Sec. 4.9].

In Fig. 7.24 we show the ΔV_T–V_p characteristics for an NMOS transistor at several different temperatures. With increasing temperature, the amplitude of the V_T shift decreases for both polarities of stress voltage and the step becomes less abrupt, approaching approaching continuity with the positive branch of the characteristic. This is consistent with our previous assertion that traps within the bandgap can follow the Fermi level at higher temperatures when interface state time constants are short but not at lower temperatures when the time constants become long relative to the stress period. At -30°C trap time constants are long and they cannot emit electrons fast enough to keep up with the Fermi level as it sweep through the bandgap. Hence, no response is seen for small negative pulses. When accumulation occurs, time constants decrease dramatically, [Sec. 4.9] and an abrupt step occurs as trapped electrons suddenly recombine with holes. An analogous argument applies to the lattice relaxation time constant.

Similar behavior can be seen in Fig. 7.25 which plots the substrate bias dependence of the ΔV_T–V_p characteristics at two different temperatures. At −30°C [Fig. 7.25(a)], ΔV_T changes very little for negative voltages until the onset of accumulation where the abrupt transition occurs. At 125°C the V_T shift initially decreases for negative voltages because interface state time constants are sufficiently short for electron emis-



(a)



(b)

Figure 7.22: Initial amplitude of threshold voltage shift versus temperature at (a) constant current (b) constant drain-source voltage.

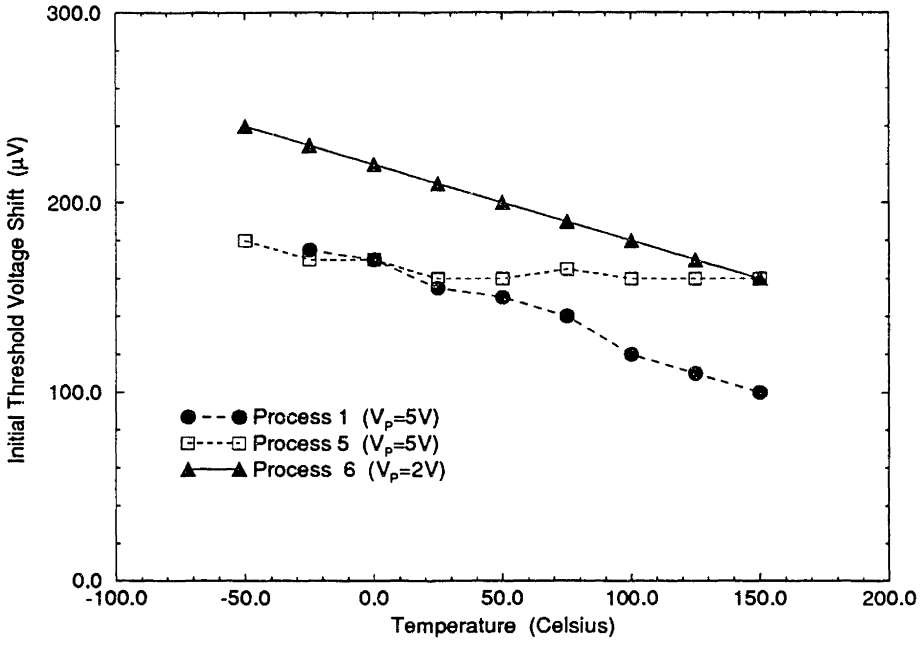


Figure 7.23: Initial threshold voltage shift versus temperature for several different processes (NMOS 20/5); $V_{GS} = 1.2V$, $V_{DS} = 500\text{mV}$, $t_s = 10\text{ms}$.

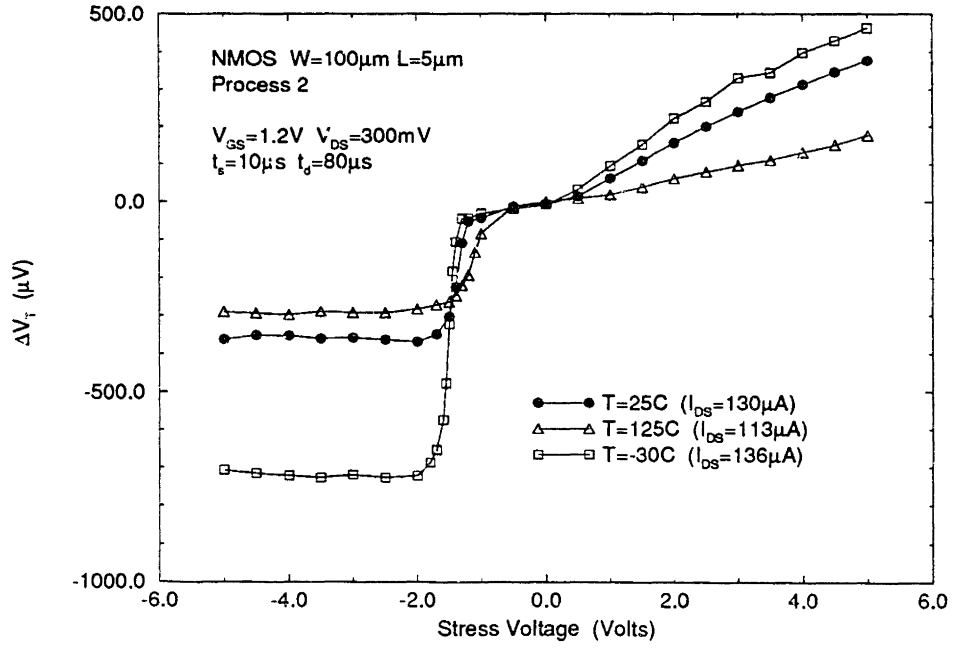
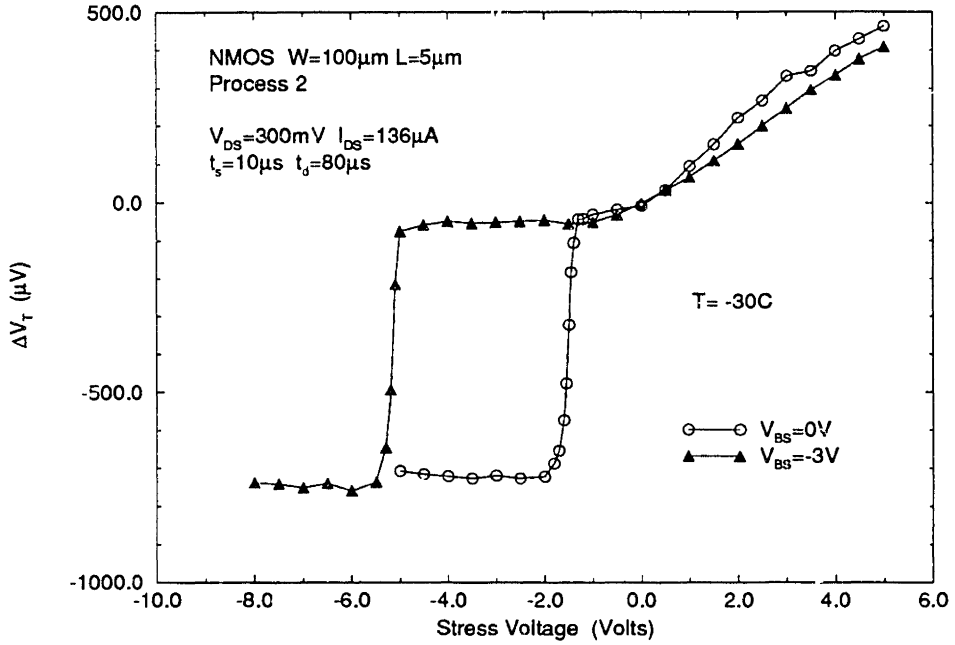
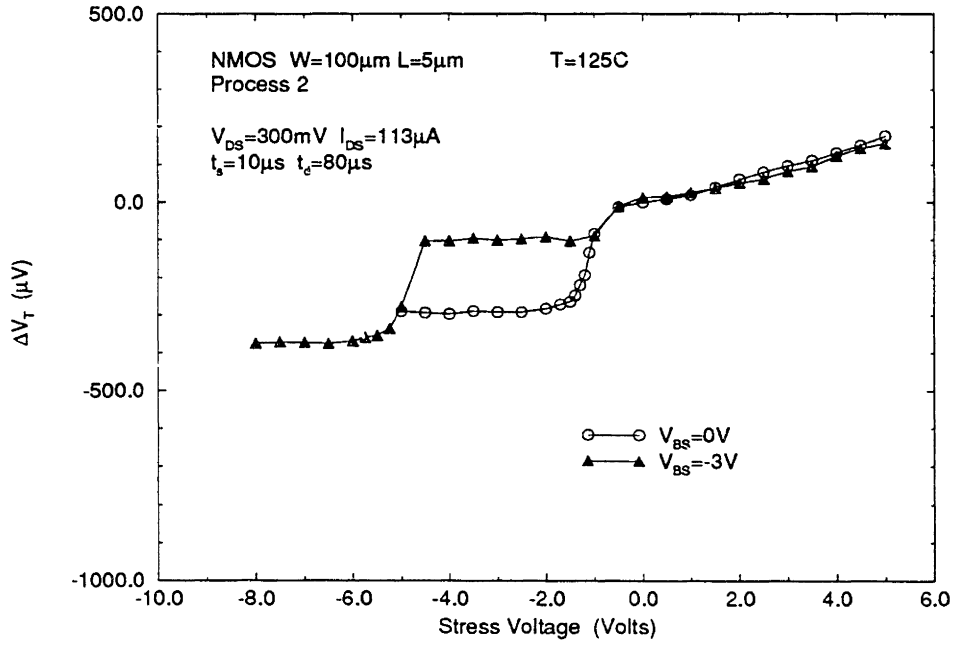


Figure 7.24: Initial threshold voltage shift versus stress voltage with temperature as a parameter.



(a)



(b)

Figure 7.25: Initial amplitude of threshold voltage shift versus stress voltage at $V_{BS} =$ 0V and $V_{BS} = -3$V at (a) -30°C and (b) 125°C.

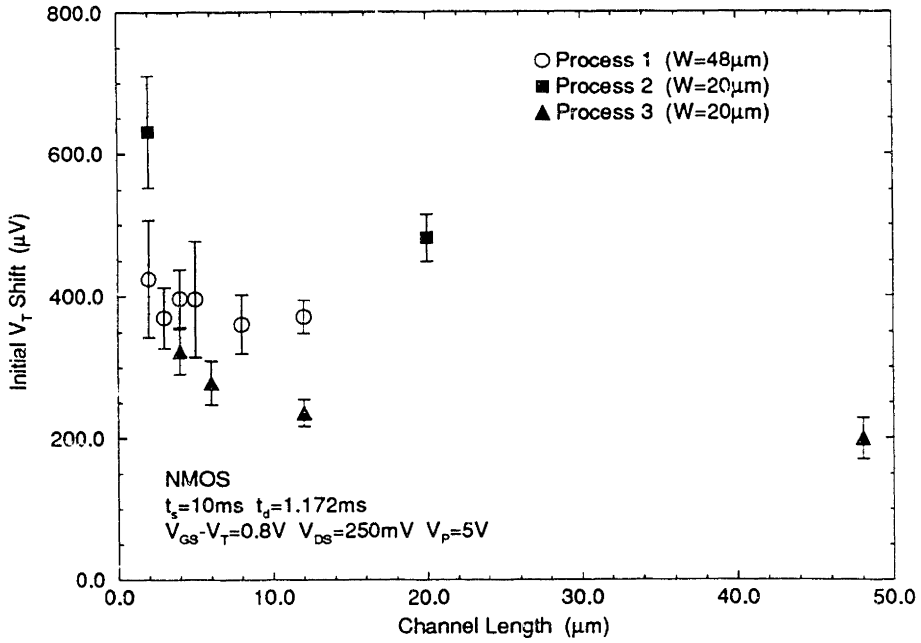


Figure 7.26: Channel length dependence

sion to the conduction band to occur as the Fermi level traverses the bandgap. Thus the $V_{BS} = -3$V curve initially tracks the $V_{BS} = 0$V curve until the Fermi level reaches energies where the trapping time constants are too long for emission to occur during the stress time. The V_T shift then reaches a plateau until the remainder of the trapped electrons are removed by recombination with holes in accumulation where, now, the accumulation point is shifted along the V_p axis by an amount given by Eq. 7.2.

Channel Width and Length

The dependence of the threshold voltage shift on channel length is shown in Fig. 7.26 for several different processes. Despite the rather large standard deviations of the measurements, as represented by the error bars, the V_T shift shows an unmistakable downward trend with increasing channel length. This is consistent with a model in which trap densities are enhanced around the periphery of the gate, as discussed in Sec. 4.6. Data on process 3 is shown in Fig. 7.27 together with a plot of Eq. 4.96. The data fits this model very well, with $\Delta V_{T0} = 200\mu$V and $\gamma_w = 1.24\mu$m.

The V_T shift also shows a slight decrease with channel width as shown in Fig. 7.28. This decrease is also described by Eq. 4.96, however too few data points are available for a meaningful fit of theory to experiment.

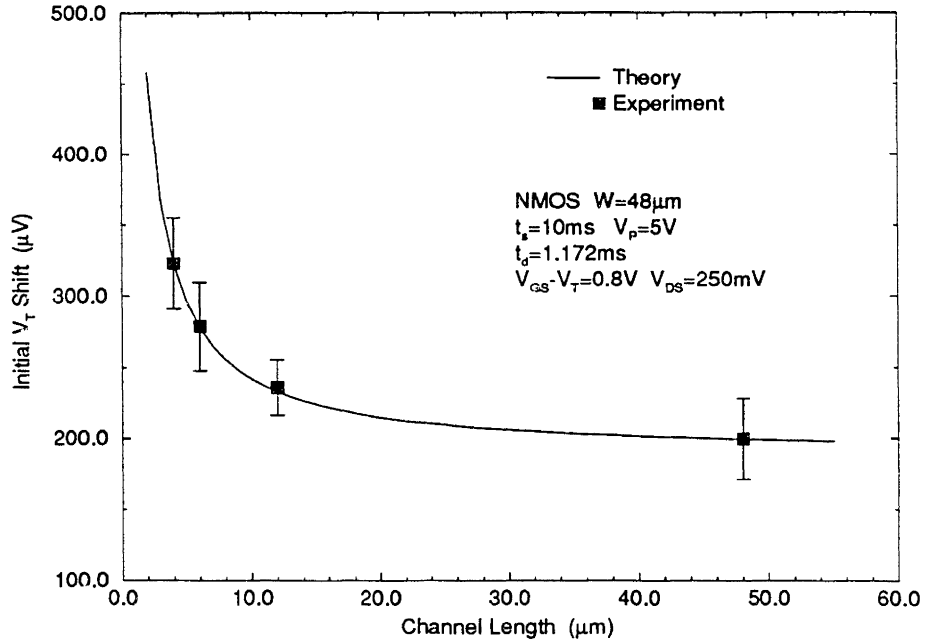


Figure 7.27: Channel length dependence (Process 3)

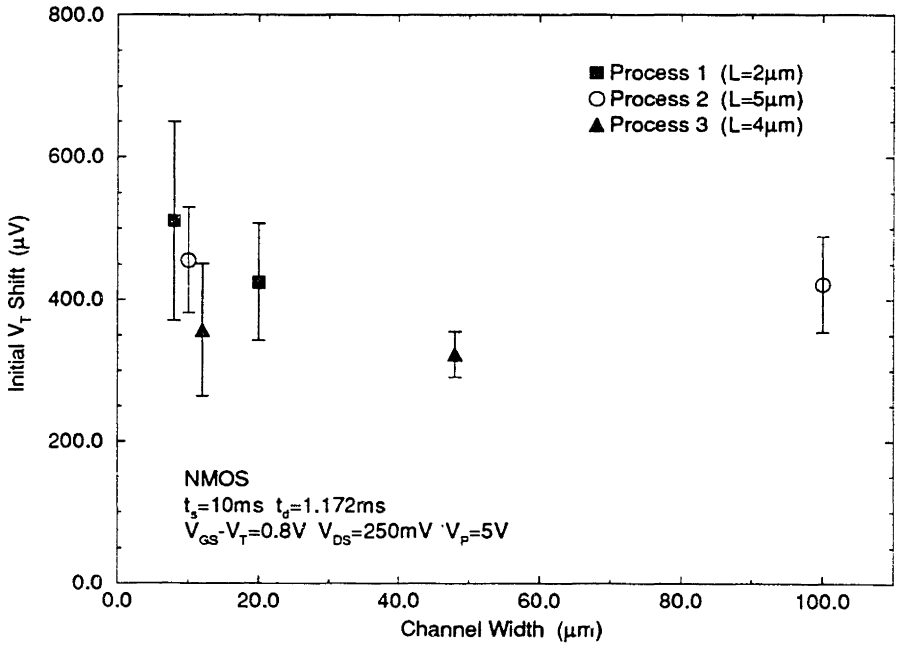


Figure 7.28: Channel width dependence

Table 7.4: NMOS (t_s = 10ms, t_d = 1ms, $V_{GS} - V_T$ = 500mV, V_{DS} = 300mV, W/L=20/5)

	$V_p = +2V$		$V_p = -2V$		$\overline{V_{GS}^2}^{1/2}$ (nV/$\sqrt{\text{Hz}}$)	
Process	ΔV_T (μV)	σ (μV)	ΔV_T (μV)	σ (μV)	Ref. [154]	Ref. [210]
(6) OXIDE	61.06	11.54	-59.82	7.59	316	158
(7) NOX	868.05	277.12	-892.05	378.94	529	—
(8) ROXNOX	843.4	82.02	-849.00	1.98	361	374

Table 7.5: PMOS (t_s = 10ms, t_d = 1ms, $V_{GS} - V_T$=-700mV, V_{DS}=300mV, W/L=20/5)

	$V_p = +2V$		$V_p = -2V$		$\overline{V_{GS}^2}^{1/2}$ (nV/$\sqrt{\text{Hz}}$)	
Process	ΔV_T (μV)	σ (μV)	ΔV_T (μV)	σ (μV)	Ref. [154]	Ref. [210]
(6) OXIDE	74.18	73.94	-12.42	5.13	251	500
(7) NOX	1564.33	217.20	-2818.00	493.08	424	—
(8) ROXNOX	1468.67	91.22	-1699.00	142.79	424	794

7.2.3 Nitrided and Reoxidized Nitrided Gate Oxides

The nitridation of silicon dioxide results in several advantageous properties which have motivated considerable research in recent years into its use as a reliable thin gate dielectric in MOSFETs. Among these advantages are a reduction in interface state generation under high electric fields [203], reduced sensitivity to radiation [204, 205] and the formation of a diffusion barrier to various dopants and contaminants [206]. However, the nitridation process also results in several disadvantageous side effects — namely, an increase in the densities of fixed positive charge [207, 208] and electron traps [96] in the dielectric and an increase in the density of interface states [207, 208]. These effects reduce device stability, degrade the inversion layer mobility and increase 1/f noise. The reoxidization of nitrided oxides has been found to overcome some of these problems by reducing the density of electron traps while preserving the advantages of nitridation [203, 209, 154].

Since threshold voltage relaxation is highly sensitive to the density of electron traps, the properties of nitrided (NOX) and reoxidized nitrided oxides (ROXNOX) under large-signal transient stress conditions are of interest as a procedure for characterizing these dielectrics, particularly if they are ever to find applications in high-accuracy analog applications. In this section we present, for the first time, measurements of threshold voltage relaxation in NMOS and PMOS transistors with thin (100Å) NOX and ROXNOX gate dielectrics.

The devices used for this study were fabricated at the MIT Microsystems Technology Laboratory by Gross et al [210] using an optimized 850°C low-pressure-furnace ROXNOX process. Figure 7.29 shows Oscilloscope photographs of the linear region

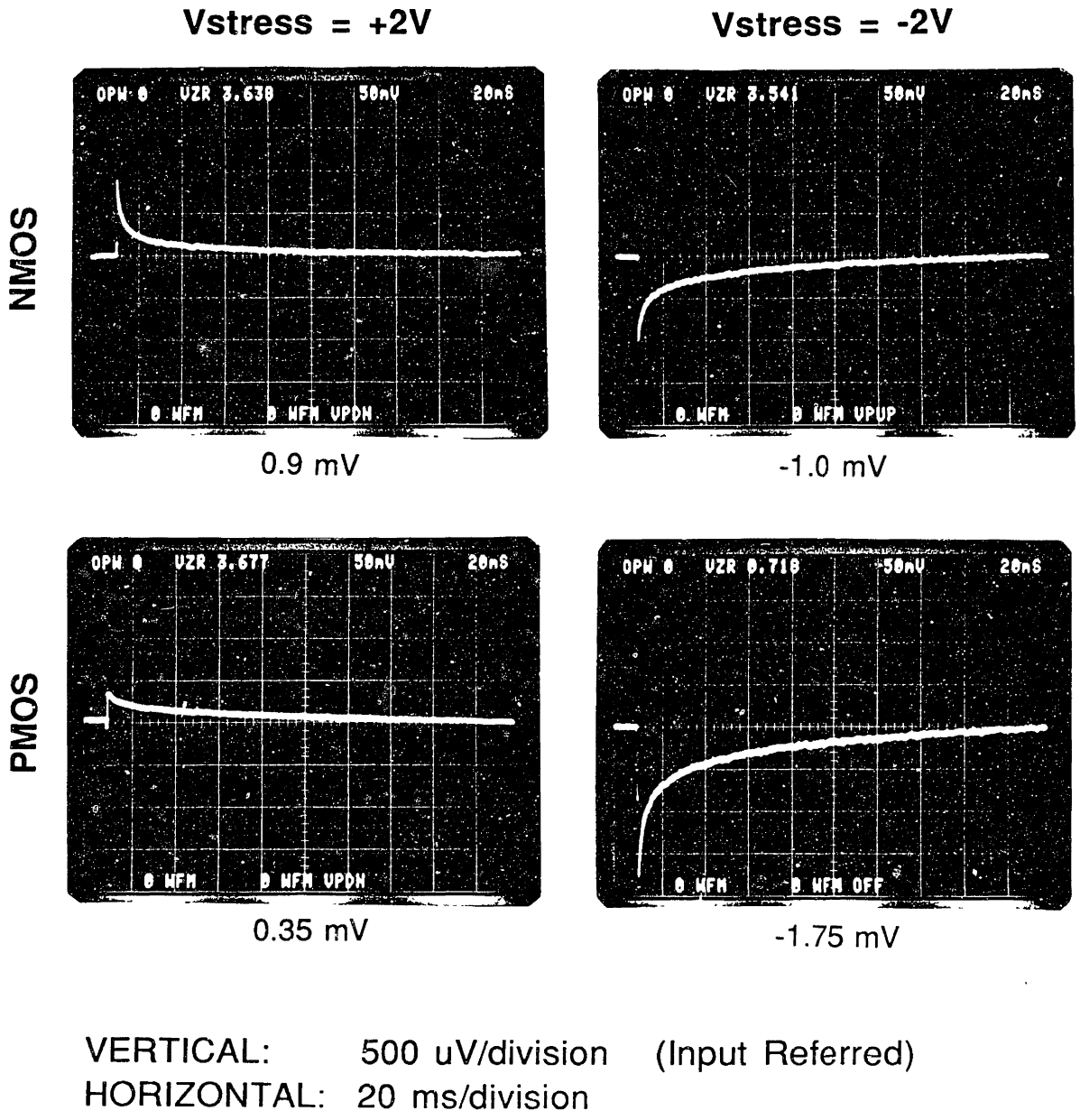
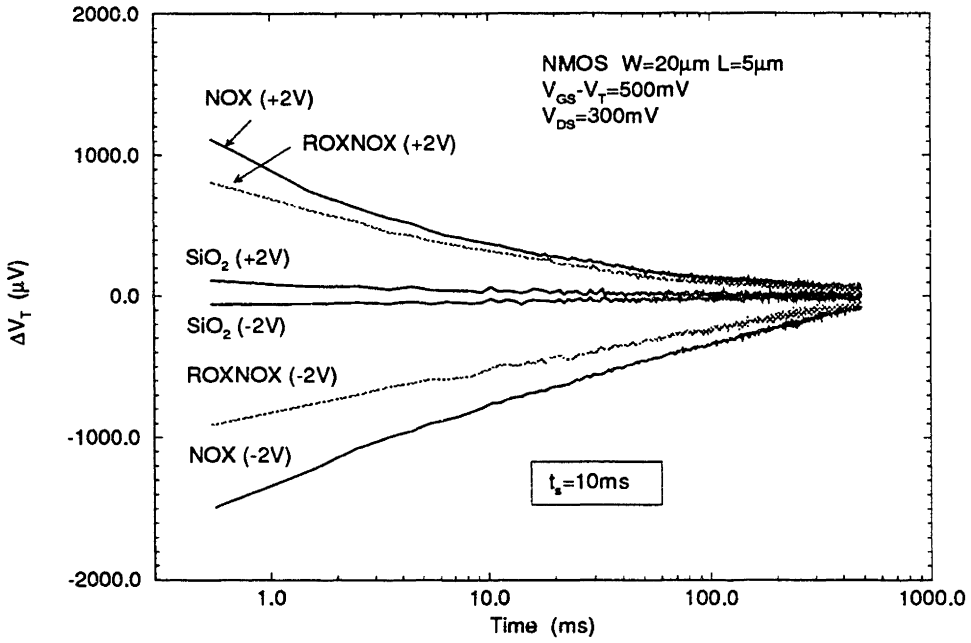
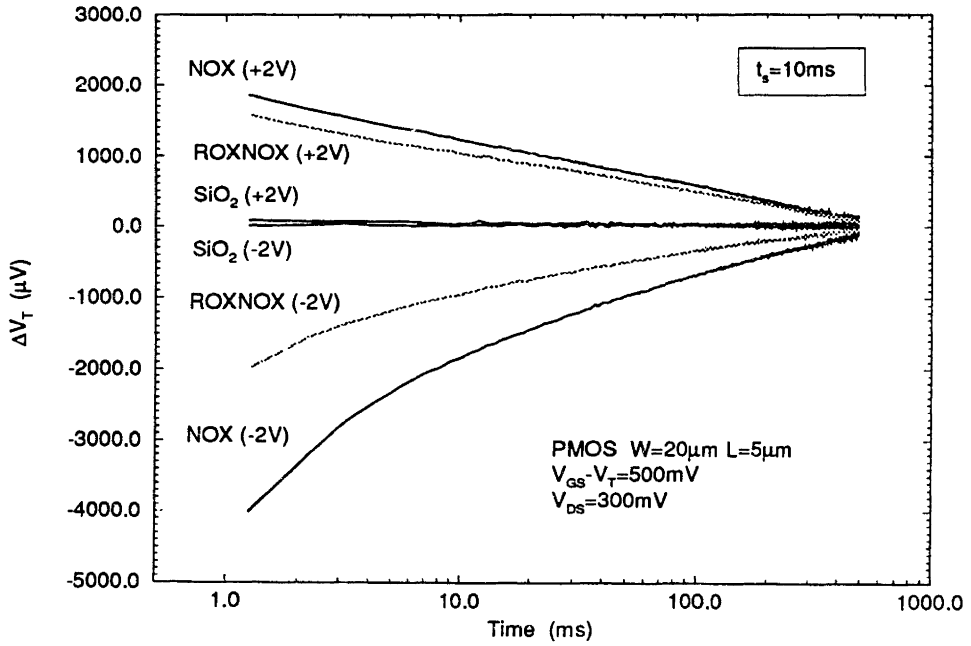


Figure 7.29: Oscilloscope photographs showing measured V_T shifts in process 7 ($W/L = 20\mu m/5\mu m$); V_{DS}=350mV, t_s = 10ms, $V_p = \pm 2$V. Vertical scale = 500μV per division (input-referred), Horizontal scale = 20ms per division.



(a)



(b)

Figure 7.30: Threshold voltage relaxation transients for 100$\AA$ oxides, nitrided oxides and reoxidized nitrided oxides; (a) NMOS (b) PMOS.

threshold voltage shifts in nitrided oxides stressed with ±2V for 10ms, using the closed-loop measurement circuit of Fig. 6.3 with a gain of 100. Table 7.4 summarizes the measurement results for 100Å conventional oxides, nitrided oxides and reoxidized nitrided oxides. Despite the very large standard deviations of the measurements, two features are immediately apparent. First, NOX and ROXNOX dielectrics exhibit V_T shifts which are a factor of 10-15 times larger than in SiO_2 control samples [process 6]. ROXNOX devices show slightly smaller threshold voltage shifts than NOX. This behavior is consistent with the larger trap densities produced by the nitridation process and the elimination of some of these traps by reoxidation [210]. Second, in contrast to conventional oxides, PMOS devices in the NOX and ROXNOX processes exhibit large threshold voltage shifts comparable to the levels observed in NMOS.

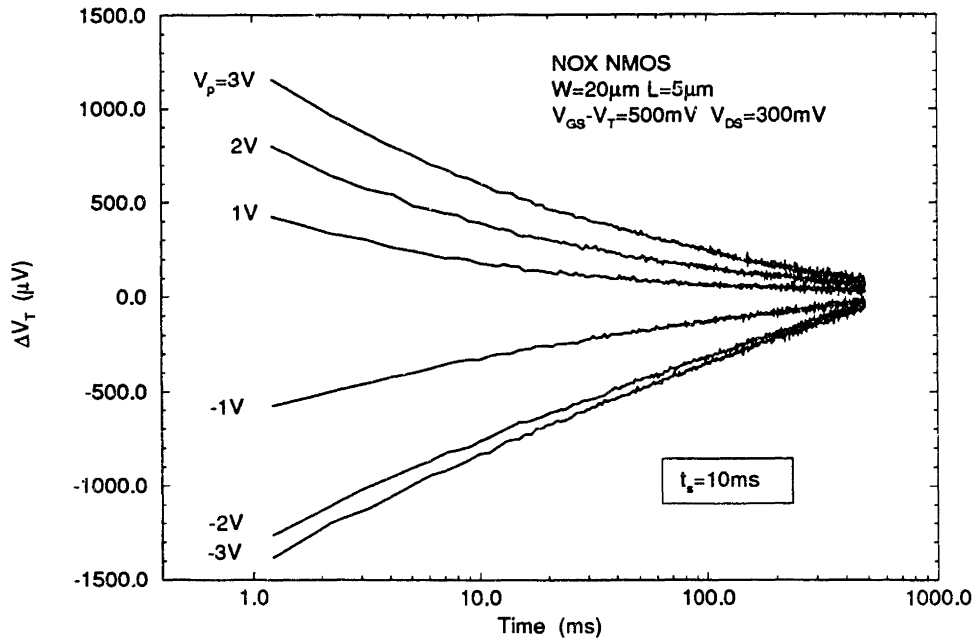
Threshold voltage relaxation transients for typical NMOS and PMOS transistors in processes 6–8 subjected to 10ms of stress at ±2V are shown in Fig. 7.30. The threshold voltage shifts are clearly greater for NOX and ROXNOX. It is also apparent that the NOX and ROXNOX transients show larger deviations from a strict logarithmic time dependence than do pure SiO_2 dielectrics. It has already been shown that the long-tail relaxation is not strictly logarithmic, even for uniform trap densities [Fig. 4.11], and this feature is amplified by the large trap densities in nitrided oxides. Nonuniformities in the trap density may contribute additional curvature.

The relaxation transients for a NOX dielectric are shown in Fig. 7.31, parameterized by stress voltage. The NMOS transistor in Fig. 7.31(a) exhibits the same general features observed in SiO_2 dielectrics in Sec. 7.2, although of larger magnitude. As in Fig. 7.7, the V_T shift increases proportional to voltage in the positive direction and shows an abrupt saturation for negative stress voltages.

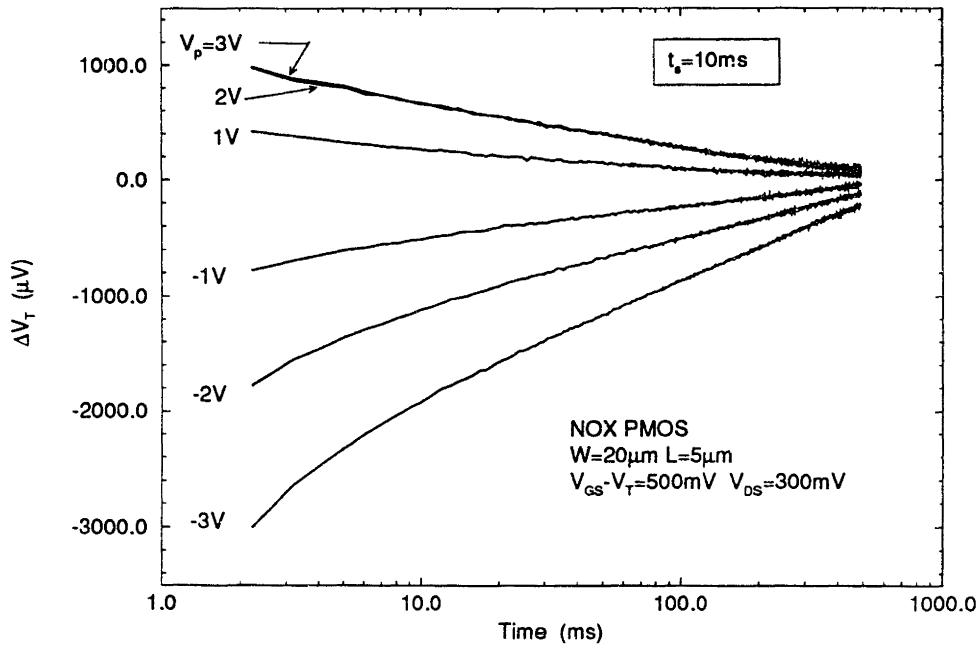
In the PMOS device shown in Fig. 7.31(b) however, we see exactly the opposite behavior, saturating for positive gate voltage pulses and increasing steadily with the magnitude of V_p in the negative direction. This indicates that holes are being trapped by oxide traps below the valence band during negative stress. The saturating behavior for positive pulses has already been described and can be explained by the rapid field emission of trapped electrons due to high local fields created by the presence of inversion layer holes when the pulse is removed.

The dependence of the V_T shift on stress time is shown in Fig. 7.32 for a NOX dielectric for several values of stress voltage. The NMOS device shows a nearly logarithmic time dependence as predicted by the tunneling theory and resembles Fig. 7.11 for conventional oxides. The PMOS device shows greater departures from a log time dependence. This can be explained by a large it/cb component [Fig. 4.13] due to a high interface state density or by to a large lr/cb component due to lattice relaxation.

The voltage dependence of the threshold voltage shifts is brought out more clearly in the ΔV_T–V_p plots shown in Figs. 7.33 for NMOS and in Fig. 7.34 PMOS devices in each of these three processes. The n-channel NOX and ROXNOX characteristics

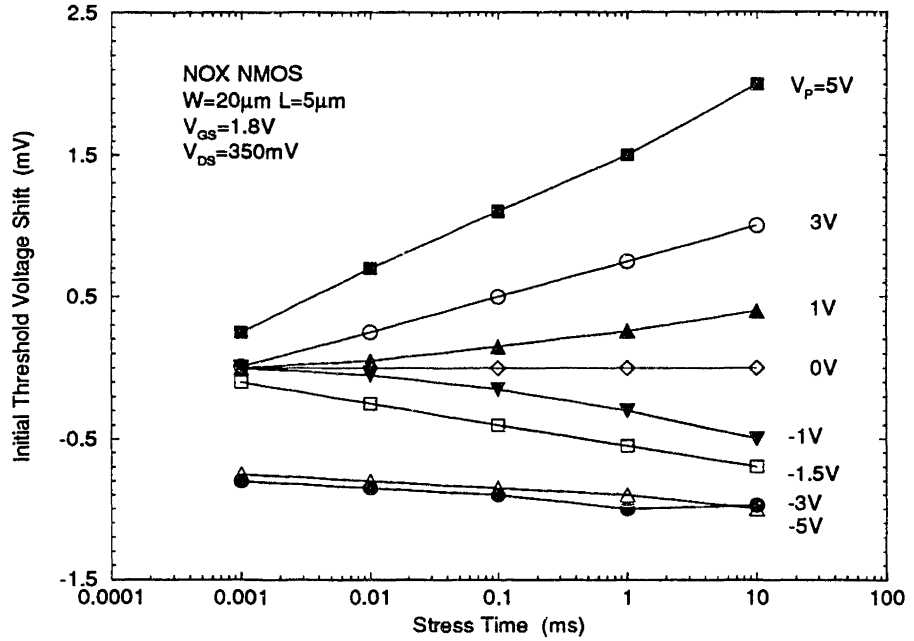


(a)

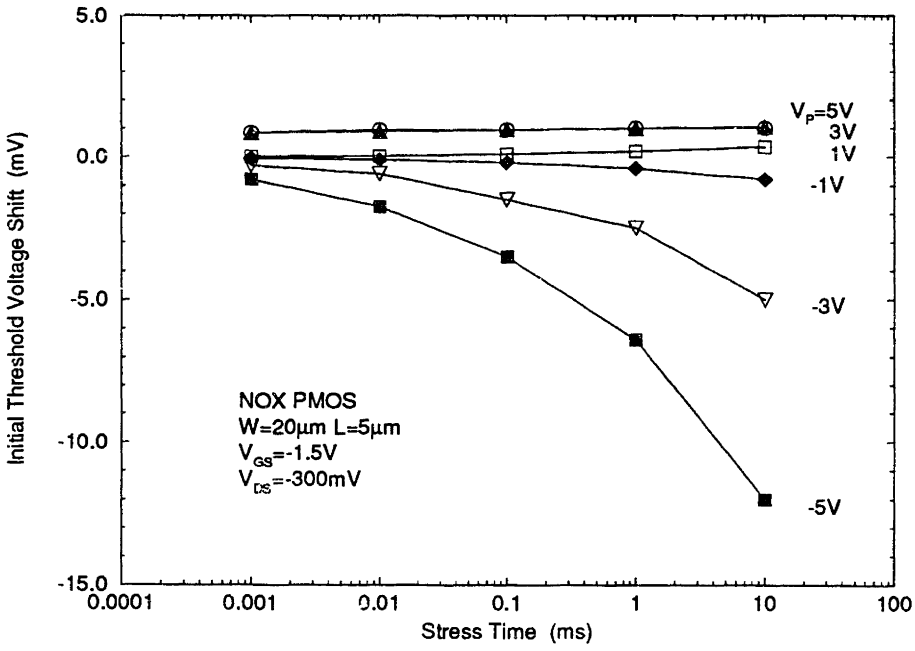


(b)

Figure 7.31: Threshold voltage relaxation transients for nitrided oxides (a) NMOS (b) PMOS



(a)



(b)

Figure 7.32: Initial threshold voltage shift vs. stress time with stress voltage as a parameter, (NOX 20/5) (a) NMOS (b) PMOS.

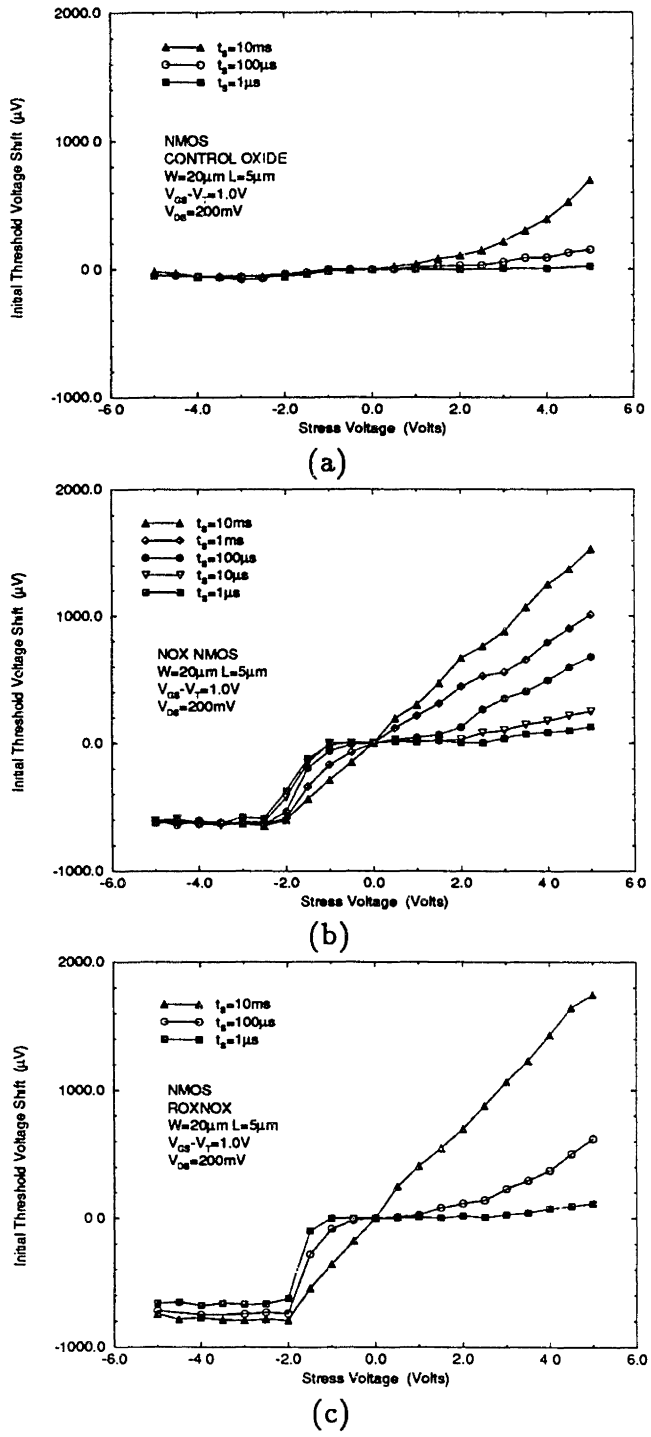


Figure 7.33: Initial threshold voltage shift versus stress voltage with stress time as a parameter (NMOS); (a) control oxide (b) NOX (c) ROXNOX.

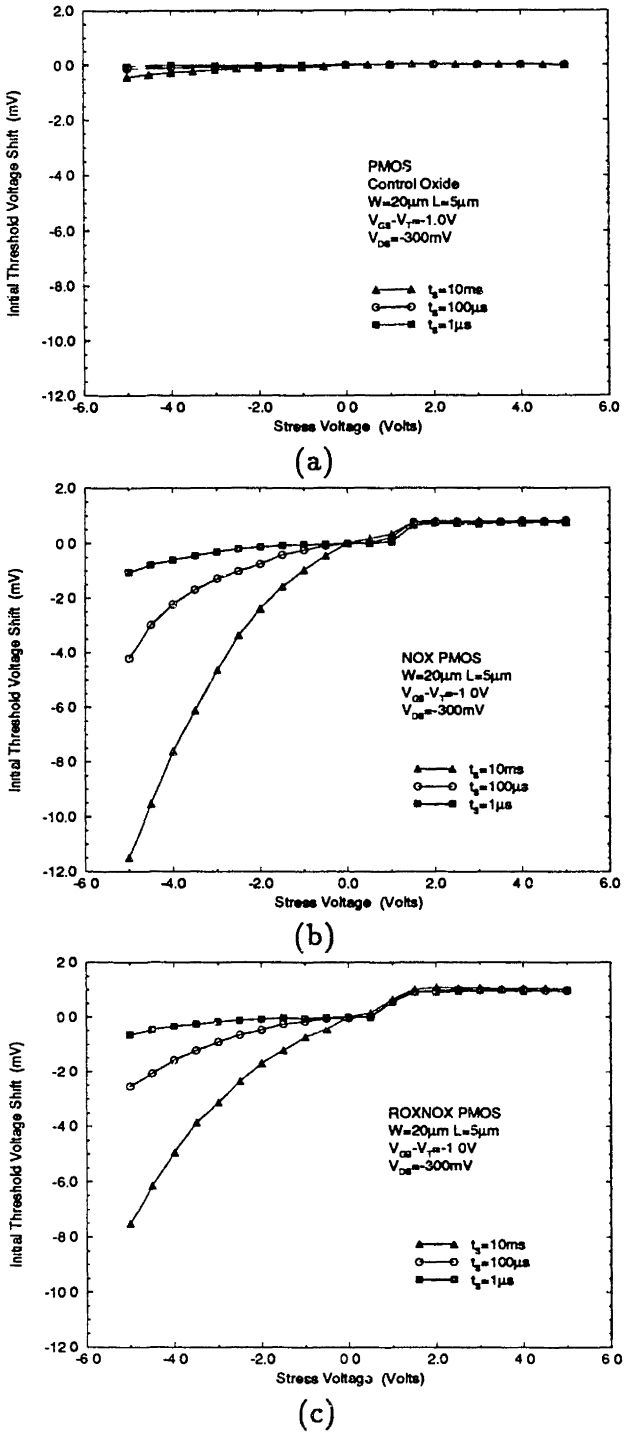


Figure 7.34: Initial threshold voltage shift versus stress voltage with stress time as a parameter (PMOS); (a) control oxide (b) NOX (c) ROXNOX.

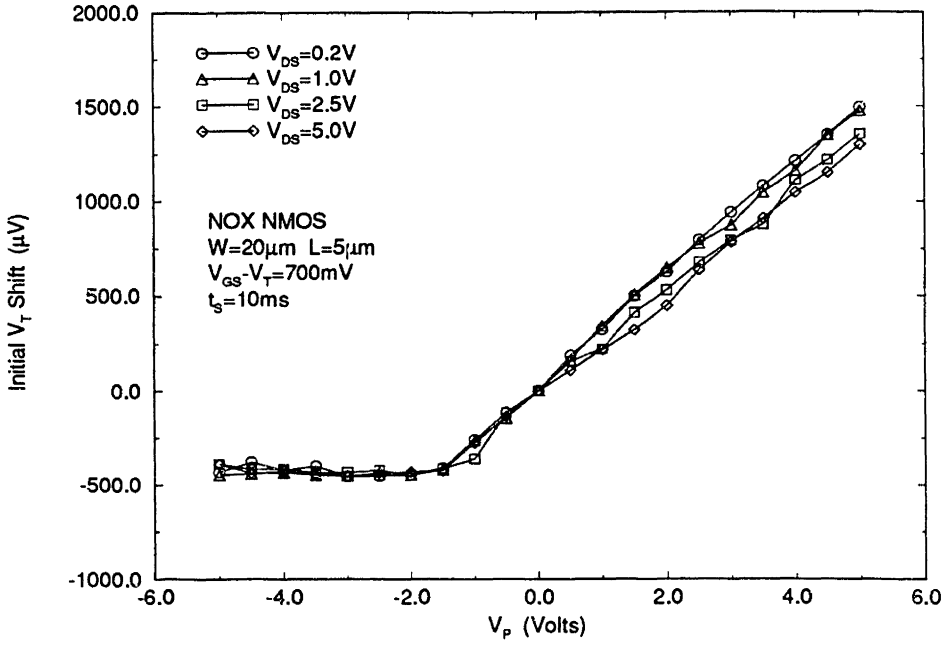


Figure 7.35: Initial threshold voltage shift vs. stress voltage with drain-source voltage as a parameter; NOX NMOS 20/5, $V_{GS} - V_T = 700\text{mV}$.

shown in Figs. 7.33 and 7.34 (b) and (c) are simply amplified versions of the SiO_2 characteristics in (a). While the behavior of the PMOS devices appears new, Fig. 7.34 suggests that the only real difference is in magnitude and that the same features are present in the voltage dependence of SiO_2 PMOS transistors at levels too small to be directly observed with our measurement circuit.

The drain voltage dependence of the V_T shifts is shown in Fig. 7.35 which plots the ΔV_T–V_p characteristics for a NOX transistor at several values of V_{DS}. It can be seen that the V_T shift is approximately constant, decreasing slightly, with increasing V_{DS} as also found in Sec. 7.2. The weak dependence of ΔV_T on drain-source voltage and hence, on power dissipation during stress, rules out self-heating as an explanation for the threshold voltage shifts.

The dependence of the threshold voltage shift on initial gate bias is shown in Fig. 7.36 for a NOX NMOS device and shows the same general features as SiO_2 [Fig. 7.17]. The explanation for this behavior is identical.

The backgate bias dependence of the threshold voltage shift is shown in Fig. 7.37. Note that the amplitude of the threshold voltage shift in the negative direction, before saturation occurs, decreases with increasing substrate bias. This may be due to the encroachment of the source and drain depletion regions into the channel. The length and width dependence data, as discussed above, suggest that there may be a significant enhancement of the trap density in the regions of the oxide adjacent to

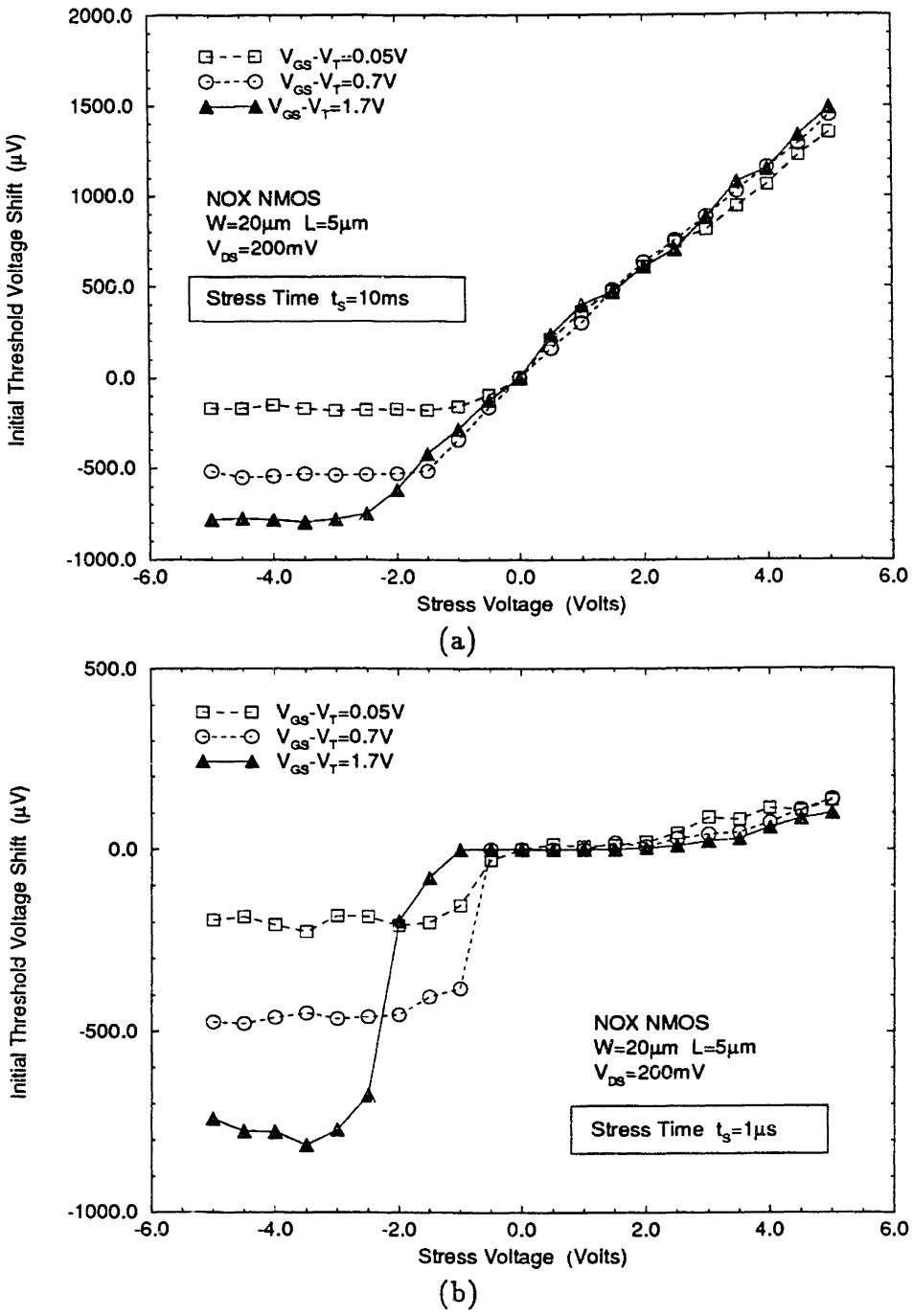


Figure 7.36: Initial threshold voltage shift vs. stress voltage with initial gate bias as a parameter, NOX NMOS 20/5, V_{DS}=200mV (a) stress time = 10ms (b) stress time = 1μs.

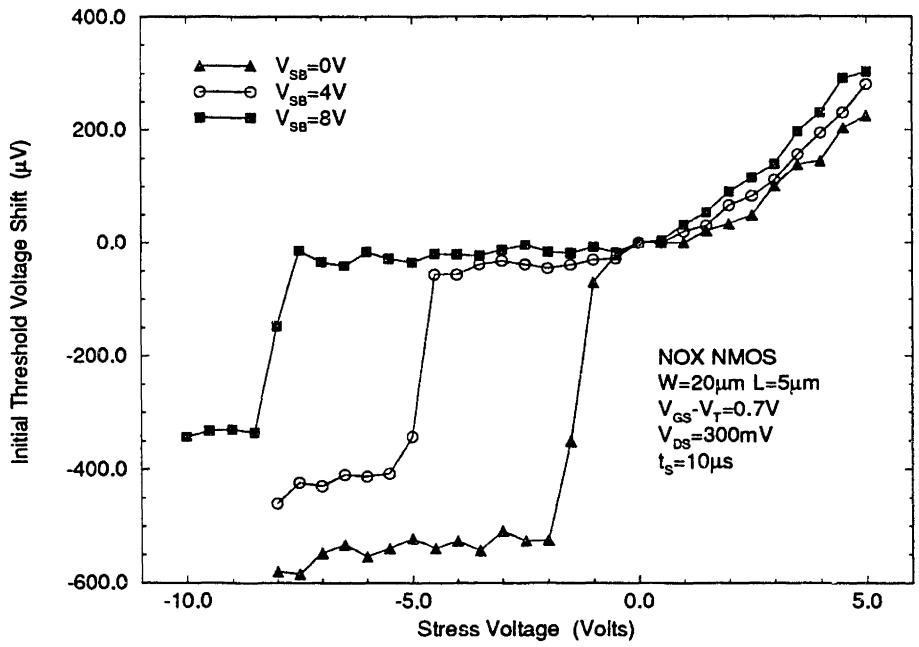


Figure 7.37: Initial threshold voltage shift vs. stress voltage with source-substrate voltage as a parameter.

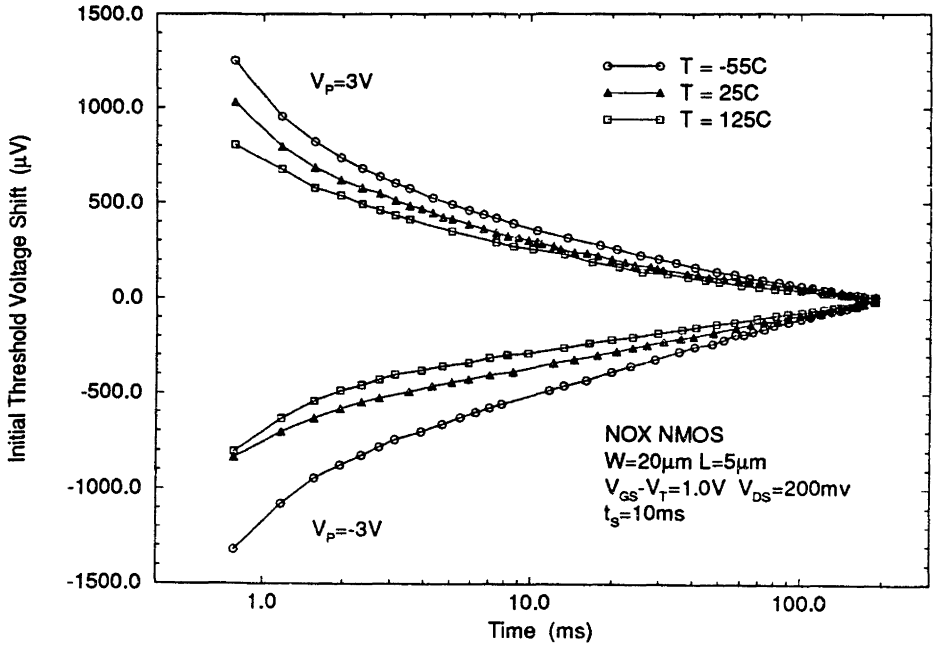


Figure 7.38: Threshold voltage relaxation transients at several different temperatures.

the source/drain implants. With increasing substrate bias, these regions of high trap density are shielded by depletion regions from majority carriers that flood the surface during stress. Note, however, that the amplitude of the threshold voltage shift in Fig. 7.19 increases slightly, rather than decreasing, with increasing substrate bias. The reason for this discrepancy is not known.

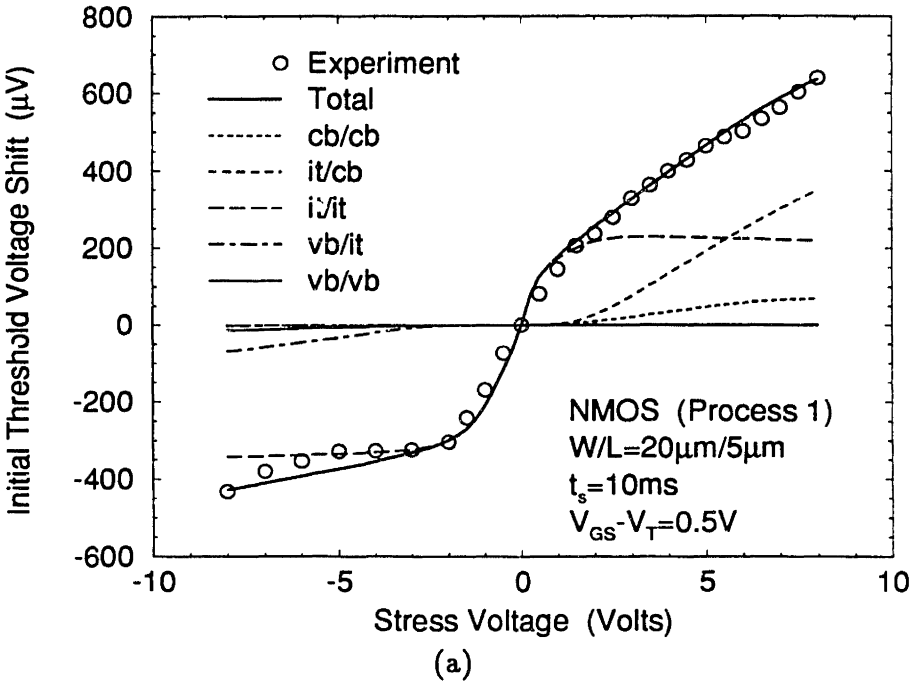
The temperature dependence of the V_T relaxation transients is shown in Fig. 7.38. The amplitude of the threshold voltage shift for both positive and negative stress voltages decreases with increasing temperature, as was found for conventional oxides.

7.2.4 Simulation Results and Extraction of Trap Densities

Theoretical considerations have led us to a description of trapping–induced threshold voltage shifts in terms of components which summarize the mechanism by which a given oxide trap fills and empties. This description is useful because it allows us to determine the dominant trap charging and discharging mechanisms and to identify their characteristic time and voltage dependences. What we observe experimentally, however, is the superposition of all of these components at once. Clever choice of bias conditions can separate out some, but not all, of these components. For example, the it/it and lr/lr components can be suppressed in strong inversion, but the it/cb, lr/cb and cb/cb components still cannot be individually distinguished. There is no experiment that allows the components to be isolated for individual observation. Thus, simulations are useful for determining the contribution of each component to the total.

Program

Due to the difficulties in obtaining exact closed-form solutions for the threshold voltage relaxation, the theory presented in Chapters 3 and 4 has been implemented in a numerical simulation program. The program computes the band bending at the initial gate-source voltage and at each pulse voltage through an iterative solution of Poisson's equation, using Fermi-Dirac statistics for the carrier concentrations [Appendix E.] Thus, the applied stress voltage need not be restricted to the Boltzmann regime. Temperature dependence enters primarily through the Fermi functions, the shift in the Fermi level and the activated factors in the interface state and lattice relaxation time constants [Eqs. 2.1, 3.99 and 3.102]. Other temperature dependent effects which are implemented, but less important over military temperature range (-55°C to 125°C) at which the measurements were conducted, are the temperature dependences of the ionized impurity concentrations [Eqs. E.18, E.19], the energy gap and the effective masses [211]. Also temperature-dependent, but not modeled in the current version of the program, is the correlated mobility term in Eq. 4.104. Mobility fluctuations may easily be included in a straightforward manner as described in Sec. 4.2.



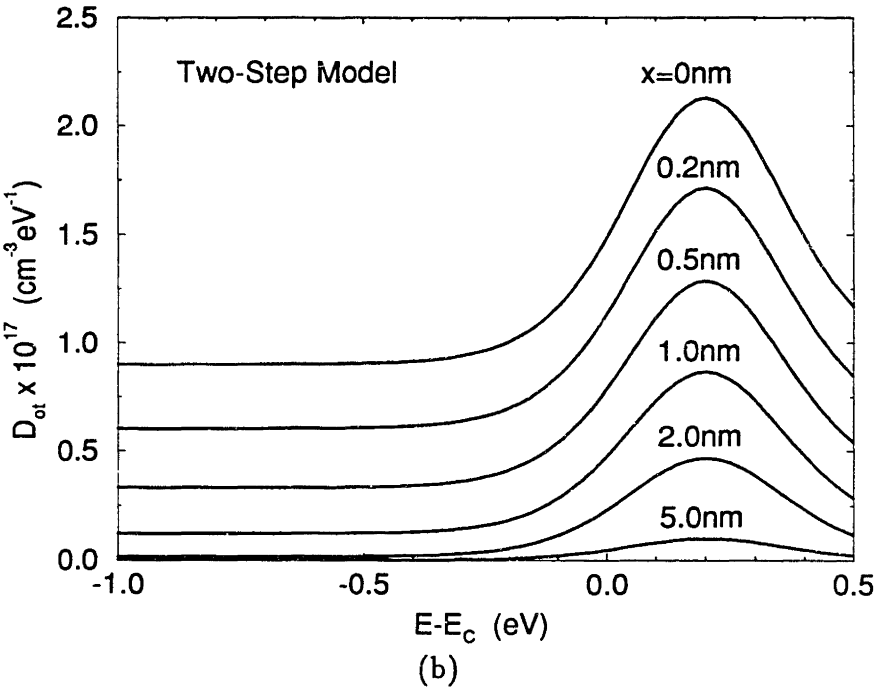
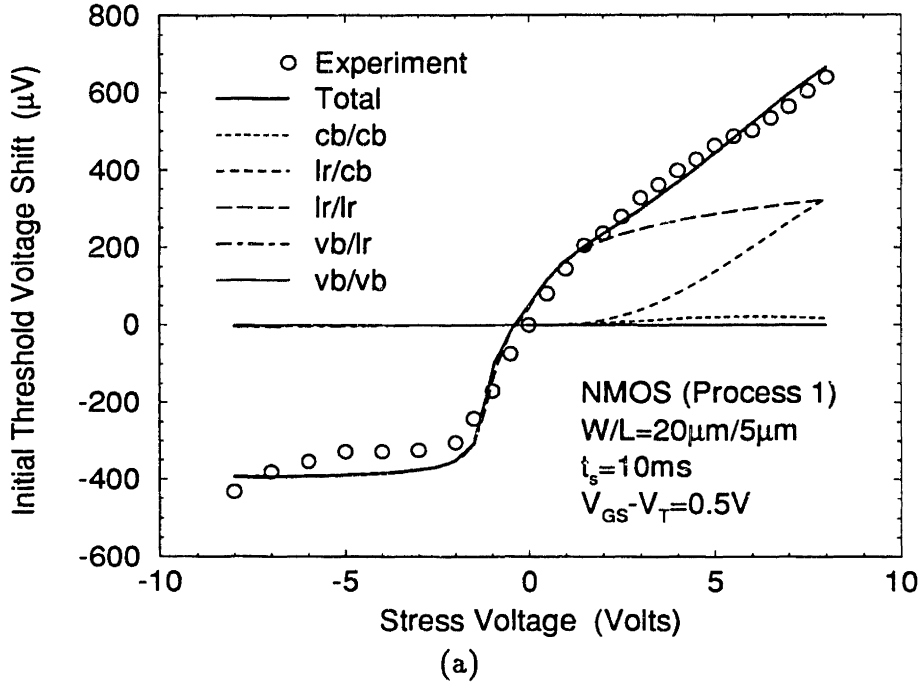
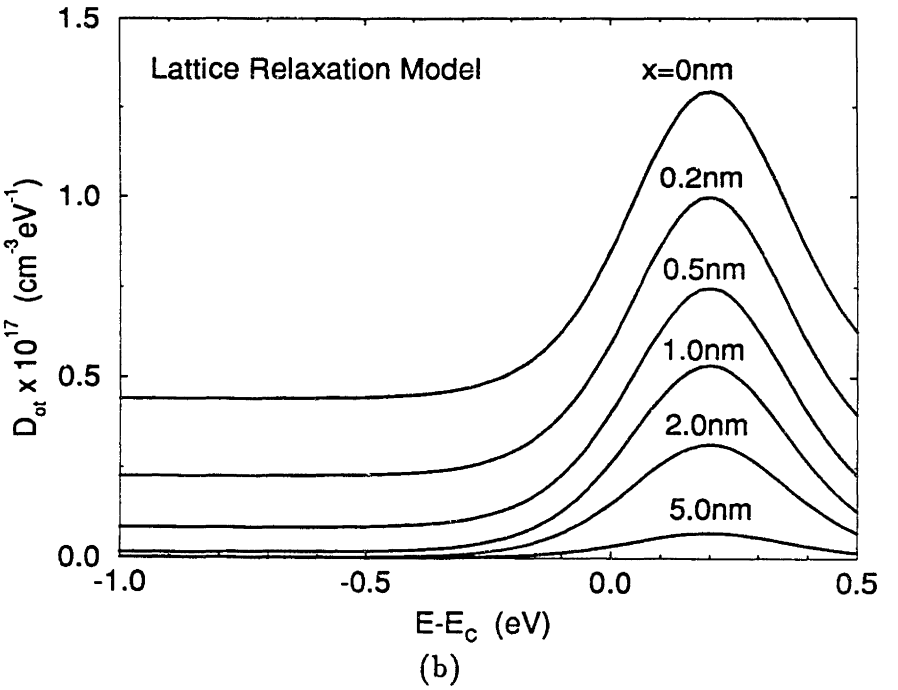


Figure 7.39: $\Delta V_T(1ms)$ vs. stress voltage; (a) Simulation using two-step model (b) Oxide trap distribution used in (a).



(a)



(b)

Figure 7.40: $\Delta V_T(t_d = 1ms)$ vs. stress voltage; (a) Simulation using lattice relaxation multiphonon emission (LRME) model (b) Oxide trap distribution used in (a).

Each of the components of the threshold voltage shift, as exemplified by Eqs. 4.41–4.43 and 4.47–4.49, is implemented as a function in the "C" programming language. The trap density $D_{ot}(E,x)$ is a user-specifiable function of energy and spatial depth. At each time point, each of the components is integrated numerically over the appropriate energy and depth limits as described in Chap. 4. The two-dimensional integrations use a 96-point Gaussian quadrature algorithm [212] for fast initial fits, and a Romberg integration [213] with a specifiable tolerance when greater precision is required. Because the trap densities which contribute to the threshold voltage relaxations resemble delta functions in space [Fig. 4.9], there is a danger, particularly at very short stress times, of these small pulses slipping undetected between the integration intervals. This problem is exacerbated in algorithms with a fixed number of intervals such as Gaussian quadrature. Fortunately, it is usually obvious when such problems arise, however the only fix is to use an algorithm with a variable number of intervals, such as trapezoidal or Romberg, with a decreased tolerance. The two-dimensional integrations therefore tend to be computationally intensive which does not permit an extensive optimization of simulation results to the data. Consequently, the simulations presented in this section represent the results of a few trial runs with physically meaningful parameter values rather than the outcome of an optimized curve-fitting procedure. The purpose is to establish that the model emulates the correct physical behavior rather than to achieve perfect fits to the data. Future versions of the simulator would benefit from enhanced computational methods to make the program efficient enough so that it could be called iteratively from a nonlinear optimization program in order to produce superior fits.

In Chapters 3 and 4, detailed theory was presented based on fundamental physics. Having provided this theoretical understanding, we now seek the simplest possible model for computational purposes which contains only those elements essential to the modeling of the experimental data. As in any model, the number of physical unknowns must be reduced to a managable number through the reduction of certain quantities to fitting parameters. We use the theory to constrain these values to physically meaningful ranges. It has been shown that the prefactor of the tunneling time constant has a relatively weak dependence on electric field, depth, and energy. It is therefore reasonable, as a first approximation, to treat these prefactors as constants, which may however differ according to the component as well as between stress and relaxation. We therefore introduce the model parameters τ_{cb1} and τ_{cb0} which represent the time constant prefactors for tunneling transitions with the conduction band during and after stress, respectively. In a similar way, τ_{it1} and τ_{it0} respresent the analogous quantities for tunneling to and from interface states. In the case of lattice relaxation, the time constant prefactor has an important dependence on the electron concentration, energy and temperature. The parameters for the LRME model are the oxide trap capture cross sections, σ_n and σ_p, and the energy barriers, ΔE_{Bc} and ΔE_{Bv}, in Eqs. 3.99, 3.35 and 3.36.

Simulations were conducted with three versions of the model, which differ only in the time constant used for traps below the silicon conduction band. The basic tunneling model uses approximation 1 of Sec. 3.5.5 in which the time constant is given by Eq. 3.102 with the SRH term τ_{it} neglected. The two-step model uses approximation 4 of Sec. 3.5.5 with the thermally activated term included in Eq. 3.102. The lattice relaxation multiphonon emission (LRME) model uses the time constant of Eq. 3.99. All of these models use the same equations for tunneling transitions with the semiconductor conduction and valence bands, as described in Sec. 3.4.1.

The WKB factors are calculated using Eq. 3.18 for tunneling from the conduction band and approximation 2 of Sec. 3.6 for the two-step process. This approximation neglects the effect of the oxide valence band on tunneling, which was shown to be valid due to the large energy barrier and effective mass of injected holes. An improved approximation can be obtained using Eq. 3.92.

The trap density $D_{ot}(E, x)$ is effectively an energy and space-dependent parameter of the model. Thus, fitting the model to the data produces, as a by-product, an estimate of the trap density, provided that the other model parameters have physically meaningful values. Since the time constant of an oxide trap depends exponentially on the spatial depth of a trap, measurements of threshold voltage shifts as a function of stress and relaxation time contain information on the spatial distribution of traps. Similarly, since the range of trap energies which contribute to the response depends on the amplitude and polarity of the applied gate voltage pulse, measurements of threshold voltage relaxation versus stress voltage contain information about the energy distribution of traps. In principle, it is therefore possible, with a set of threshold voltage relaxation measurements which is in some sense "complete", to extract oxide trap spatial and energy distributions. While we have not yet perfected this technique, or devised a detailed strategy, we shall indicate the basic approach and report on some estimated trap densities below.

Results

Because of their larger threshold voltage shifts, NMOS devices were chosen for the simulations in order to minimize uncertainties and experimental errors in the data. Experimental threshold voltage shifts as a function of stress voltage are shown in Fig. 7.39, together with simulation results using the two-step model. The decomposition of the total threshold voltage shift into its various components is also shown. As expected from the approximate models of Sec. 4.4, the it/it component changes rapidly for stress voltages which drive the electron quasi-Fermi level through the bandgap and saturates at large positive and negative voltages when the Fermi level becomes pinned at the band edges. Under the stated bias conditions, the $c\dot{b}/cb$ and it/cb components increase proportional to voltage in the positive direction and contribute negligibly for negative voltages. The threshold voltage shift in the negative direction is overwhelmingly dominated by the it/it component, with a small contri-

bution from the vb/vb component. This justifies our earlier claim [Sec. 4.4] that the NMOS threshold voltage shifts are well described by the the three dominant components – it/it, it/cb and cb/cb. The total simulated threshold voltage shift shows excellent agreement with the data.

We must excercise caution, however, in interpreting the simulation results for large negative voltages, since we have seen that accumulation results in the capture of majority carriers which are rapidly detrapped by field emission when the surface returns to inversion. This effect has not been modeled in the simulation program. Consequently, no attempt is made to extract trap densities at energies below the valence band in NMOS or above the conduction band in PMOS.

The trap distribution which was used to obtain the fit in Fig. 7.39(a) is given by the expression

$$D_{ot}(E,x) = D_0 e^{-x/x_o} + D_1 \left[\frac{1}{1+e^{(E-E_L)/\Delta E_L}}\right]\left[1 - \frac{1}{1+e^{(E-E_H)/\Delta E_H}}\right] e^{-x/x_1} \quad (7.4)$$

with parameters as specified in the first column of Table 7.6. σ_{no} and σ_{po} are the interface state electron and hole capture cross sections respectively. Contours of this distribution at constant x are plotted in Fig. 7.39(b) as a function of energy above the silicon conduction band edge. The energy distribution shows a broad peak at about 200meV above the conduction band, falling to a lower, more uniform density in the bandgap. The energy distribution appears to be continuous, which is reasonable since the long-range disorder of amorphous SiO_2 would cause substantial broadening of discrete trap levels.

A single time measurement such as this is not sufficient to uniquely determine the oxide trap spatial density, which contains more degrees of freedom than the measurement excercises. There may also be finer structure in the energy distribution which is smoothed out by our large-signal measurements. A more thorough characterization should include a family of curves like that of Fig. 7.39(a) over a range of stress and relaxation times. We do not attempt to extract any more detailed information than the degrees of freedom in our measurement permit. The decrease in the distribution at high energies is primarily a mathematical convenience, since these energies are not actually probed by the measurement and, in most cases, a trap density which levels off after the peak ($E_H \rightarrow \infty$) has been found to provide an equally satisfactory fit. Nevertheless, all of the NMOS samples which were simulated required a trap distribution of the general form of Eq. 7.4, with slight variations in the energetic position (E_L) and width (ΔE_L) of the peak, and the densities (D_0, D_1).

The two-step model was replaced with the LRME model in Fig. 7.40(a). It is evident that this model is also capable fitting the experimental data, with the parameters specified in column 2 of Table 7.6. The resulting oxide trap distribution is shown in Fig. 7.40(b) and, not surprisingly, differs somewhat from that of Fig. 7.39(b). The extracted trap distribution depends somewhat on the model assumed for the filling of traps below the Fermi level, although this may result in part from errors in the values

Table 7.6: Model Parameters used in Simulations

Parameter	Figure				
	7.39	7.40	7.41–7.45	7.46	7.47–7.48
Process	1	1	2	5	7
Model	2-step	LRME	2-step/LRME	basic	basic
D_0 (cm^{-3})	7.0×10^{16}	4.4×10^{16}	6.0×10^{16}	2.5×10^{16}	1.0×10^{17}
D_1 (cm^{-3})	2.3×10^{17}	1.6×10^{17}	2.9×10^{17}	2.5×10^{17}	3.4×10^{18}
E_L (eV)	0.1	0.1	0.1	0.0	0.0
ΔE_L (eV)	0.1	0.1	0.1	0.025	0.025
E_H (eV)	0.3	0.3	0.3	∞	∞
ΔE_H (eV)	0.1	0.1	0.1		
x_o ($\mathring{A}$)	5	3.0	5.0	5.0	5.0
x_1 ($\mathring{A}$)	20.0	20.0	20.0	50.0	50.0
τ_{cb0} (s)	1.0×10^{-10}	1.0×10^{-10}	1.0×10^{-8}	1.0×10^{-10}	1.0×10^{-8}
τ_{cb1} (s)	1.0×10^{-10}	1.0×10^{-10}	1.0×10^{-8}	1.0×10^{-9}	1.0×10^{-9}
τ_{it0} (s)	1.0×10^{-10}	1.0×10^{-10}	1.0×10^{-8}	1.0×10^{-8}	1.0×10^{-7}
τ_{it1} (s)	1.0×10^{-10}	1.0×10^{-10}	1.0×10^{-8}	1.0×10^{-9}	1.0×10^{-8}
σ_{no} (cm^2)	1.0×10^{-15}		1.0×10^{-15}		
σ_{po} (cm^2)	1.0×10^{-15}		1.0×10^{-15}		
ΔE_{Bc} (eV)		0.0	variable		
ΔE_{Bv} (eV)		0.0	variable		
σ_n (cm^2)		1.0×10^{-17}	variable		
σ_p (cm^2)		1.0×10^{-17}	variable		

for some of the other parameters. Nevertheless, this emphasizes the importance of identifying the dominant mechanism if this technique is to be used for the accurate extraction of oxide trap densities. The fact that either model is capable of fitting the data makes an experimental determination of the dominant mechanism difficult. We also note, from Eq. 4.104, that the correlated mobility term will affect the magnitude, but not the shape, of the extracted trap distribution. These questions regarding the relative importance of various mechanisms must be precisely quantified before this method can be used to reliably extract oxide trap distributions.

Relaxation transients for an NMOS transistor stressed at $V_p = 5V$ for 10ms are shown in Fig. 7.41(a) at several different temperatures. Simulations using the two-step model are also shown and successfully predict the decrease of the threshold voltage shift with increasing temperature. The deviations from a strict logarithmic time decay are also well modeled. The decomposition of the total threshold voltage shift at T=−30°C into its components is shown in Fig. 7.41(b). An increasing trap density above the conduction band [Table 7.6] is necessary in order to obtain the correct temperature dependence. This is consistent with the discussion of Sec. 4.5, in which it was found that the temperature dependence for positive pulses was governed, in the two-step model, by (1) the movement of the Fermi probe to lower temperatures with increasing temperature, (2) a trap density increasing in energy above the conduction band.

Since the LRME and two-step models have very different temperature dependences in strong inversion [Sec. 3.8], it is interesting to compare their performance in this regime. The temperature dependence of the threshold voltage shift in Process 2 after 10ms of stress with $V_p = 5$V, followed by a 0.57ms delay is shown in Fig. 7.42. Figure 7.42(a) shows simulations using the two step-model with the same parameter set as used in Fig. 7.41. The model provides an excellent fit to the data. The LRME model was used with the same parameter set but with the capture cross section $\sigma_n = \sigma_p$ in the LRME time constant [Eq. 3.99] adjusted such that both models give the same ΔV_T at room temperature. The LRME model is also capable of fitting the data, as shown in Fig. 7.42(b), but only when the barrier height ΔE_B is zero. The figure shows that the exponential decrease in the time constant with increasing temperature causes the threshold voltage shift to increase rapidly with temperature, contrary to the data. This effect which becomes more pronounced with increasing values of ΔE_B. Since reported values for ΔE_B range from 100-600 meV [116], the temperature dependence does not lend strong support to a lattice relaxation mechanisms for the threshold voltage shifts. In Sec. 4.5, it was shown that, neglecting correlated mobility fluctuations, the temperature dependence for negative pulse voltages was dominated by the activated nature of the two-step and LRME time constants for deep traps when the free carrier concentrations are low. This is investigated in the ΔV_T versus stress voltage characteristics shown in Fig. 7.43. Simulations using the two-step model are shown, with the same parameters as Fig. 7.41, and correctly model the abrupt

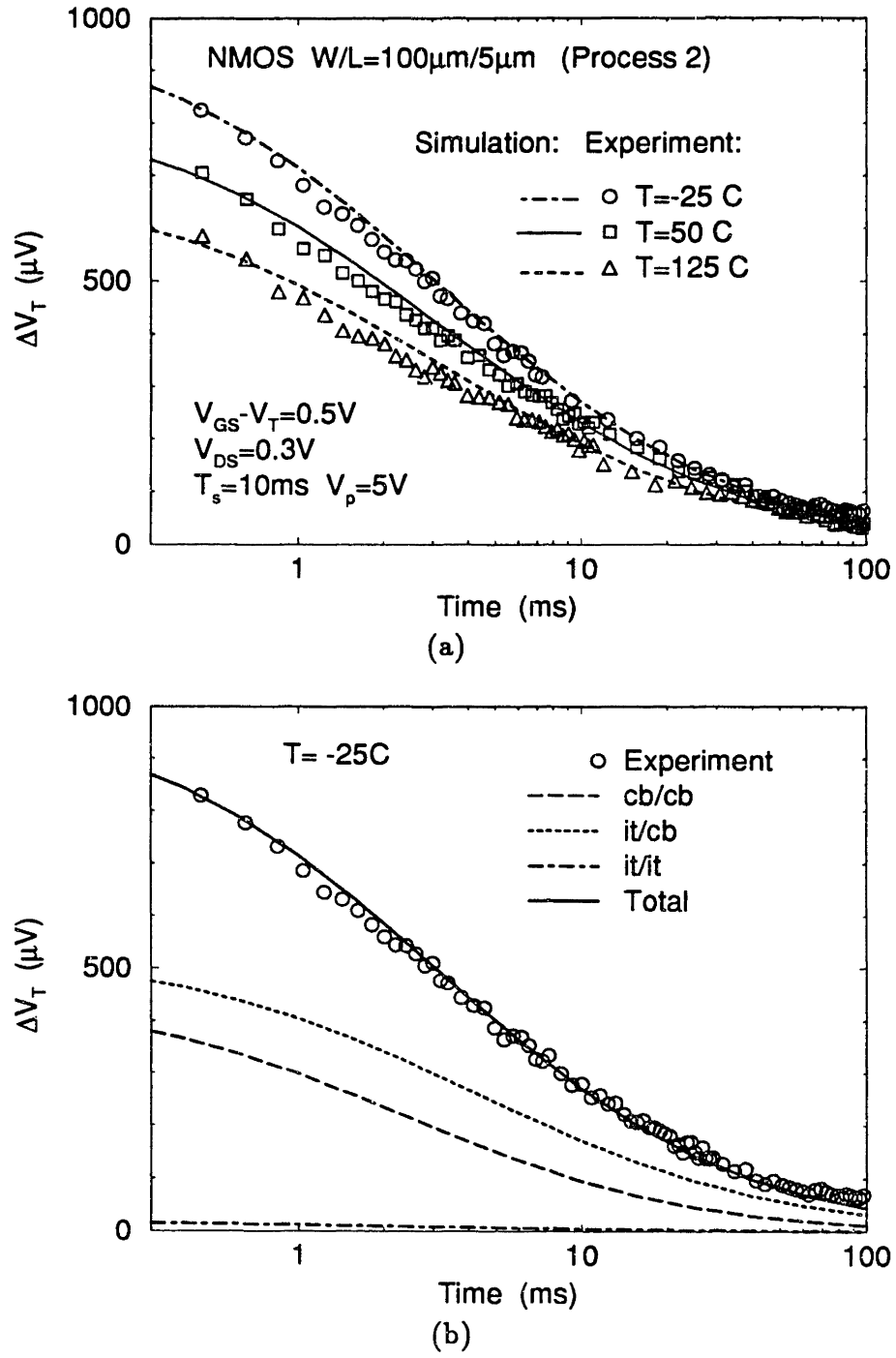


Figure 7.41: (a) Relaxation transients at three temperatures (b) Decomposition by component.

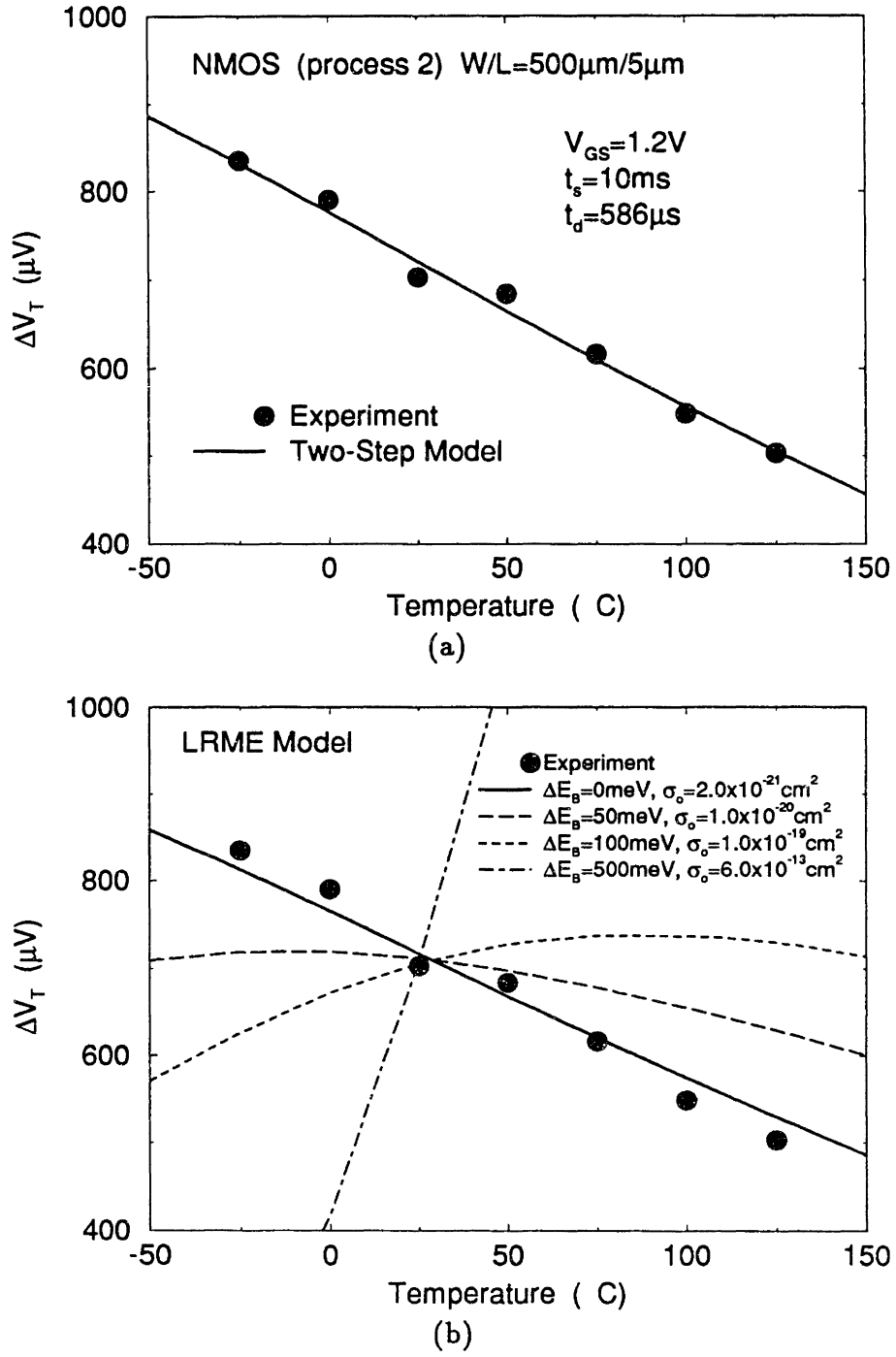


Figure 7.42: Temperature dependence in strong inversion (a) Two-step model, (b) Lattice relaxation MPE model.

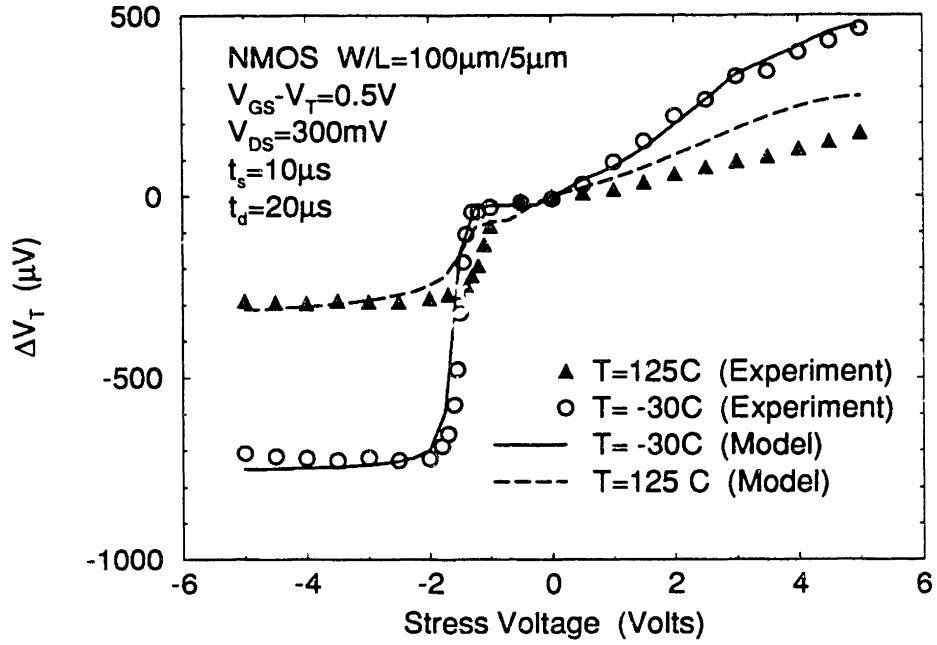


Figure 7.43: Simulated and measured $\Delta V_T(t_d = 20\mu s)$ vs. V_p characteristics at two different temperatures.

step in the characteristic at low temperatures as the Fermi level passes through the bandgap. At higher temperatures the saturation level decreases and the transition broadens. This behavior, which has already been explained in Secs 4.5 and 7.2.2, is well modeled.

Figure 7.44 compares the various models; the basic tunneling model [Eq. 3.102 with $\tau_{it} = 0$], the two-step model [Eq. 3.102] and the LRME model [Eq. 3.99], on the data of Fig. 7.43. The trap densities in all cases were identical, as specified in Table 7.6. These trap densities were obtained by fitting the data to the two-step model at $T = -30°C$ and consequently, the LRME model with the same densities does not produce an optimal fit. It is assumed that $\Delta E_B = 0$ in the LRME model for the reasons described above. At high temperatures (T=125°C), the time constants are all nearly temperature-independent and all the simulation results show the same qualitative behavior [Fig. 7.44](a). As the temperature is decreased to $-30°C$ [Fig. 7.44(b)], the time constant in the basic model remains unchanged and the threshold voltage shift is underestimated. The two-step and LRME models, both of which contain thermally-activated factors of the same form, show an abrupt transition to a larger saturation level.

It can be concluded, at this point, that both the LRME and two-step processes are plausible mechanisms for the filling of traps below the conduction band energy, provided that $\Delta E_B = 0$ in the former. Under this condition, it would be difficult to

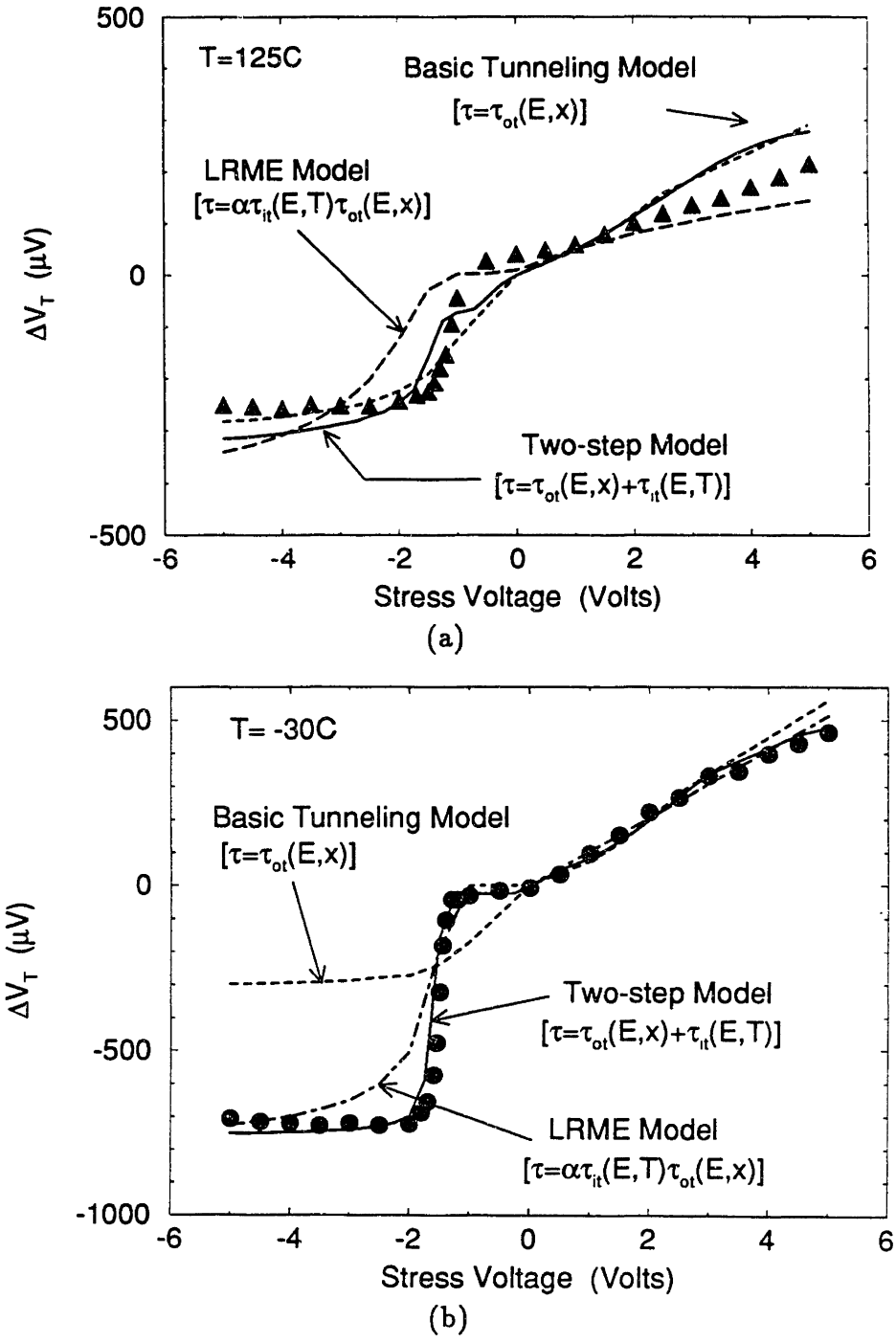


Figure 7.44: Comparison of models to temperature dependence of ΔV_T vs. V_p; (a) T= 125°C (b) T= −30°C.

distinguish these two mechanisms using our measurements. However other techniques, e.g. 1/f noise [116], do not find the barrier height to be zero. While our temperature measurements appear to favor a two-step process, it is possible that unmodeled effects, e.g. mobility fluctuations, may offset the positive temperature coefficient of ΔV_T due to lattice relaxation and bring the theory back in line with the data. Therefore, it is not possible, on the basis of these measurements, to confirm which of the two mechanisms dominates and further research in this area is needed.

The dependence of ΔV_T on initial gate bias is shown in Fig. 7.45, using data from Fig. 7.16. The model, using the parameter set of column 3 in Table 7.6, accurately predicts the decrease in ΔV_T with increasing initial level of inversion. The reason for the decrease was discussed in Sec. 7.2.2 and is indicated schematically in Fig. 7.45(b). At low V_{GS}, a large concentration of traps above the initial Fermi level are empty and can be filled during stress, contributing to the theshold voltage shift. Increasing the initial gate bias tends to pin the Fermi level against the conduction band, restricting its movement, and band bending in the oxide pulls the peak of the trap distribution below the Fermi level where the traps are always filled. Positive pulses are then less effective in filling oxide traps. The reverse argument holds for negative pulses.

A family of ΔV_T curves for different stress voltages are plotted in Fig. 7.46 versus stress time. The parameters for the simulation are given in column 4 of Table 7.6. Considering that no curve fitting was done, the model shows good agreement with the data over voltage and time, including the abrupt transition for negative pulse voltages.

Simulations have also been conducted for NOX NMOS transistors, using the two-step model with a trap distribution having the same form as Eq. 7.4 with parameters as given in column 5 of Table 7.6. Figure 7.47 shows a simulated relaxation transient, indicating the decomposition by component. The largest contributions are from the *cb/cb* component which shows a log(time) dependence and the *it/cb* component which contributes to the curvature through the $\log^2$(time) dependence (Eq. 4.78). The *it/it* contribution is smaller and shows a strong log(time) dependence. All of the other components are negligibly small at these stress levels.

Threshold voltage shifts on the same device are plotted in Fig. 7.48 as a function of stress voltage for a stress time of 10ms. In Fig. 7.48(a), the components of the threshold voltage shift are shown and the total shift successfully models the data, saturating for negative voltages and increasing approximately linearly in strong inversion. In Fig. 7.48(b), the same experiment is conducted at several values of stress time. While the simulation is not perfect, it exhibits good qualitative agreement with the data. The simulation in this case used the basic tunneling model, with a temperature-independent time constant, and therefore the modeled step is less abrupt than the data shows. This problem can be remedied by adding the thermally activated term τ_{it} to the tunneling time constant [Eq. 3.102], as shown in Fig. 7.43.

As expected from their larger threshold voltage shifts, the extracted trap densities

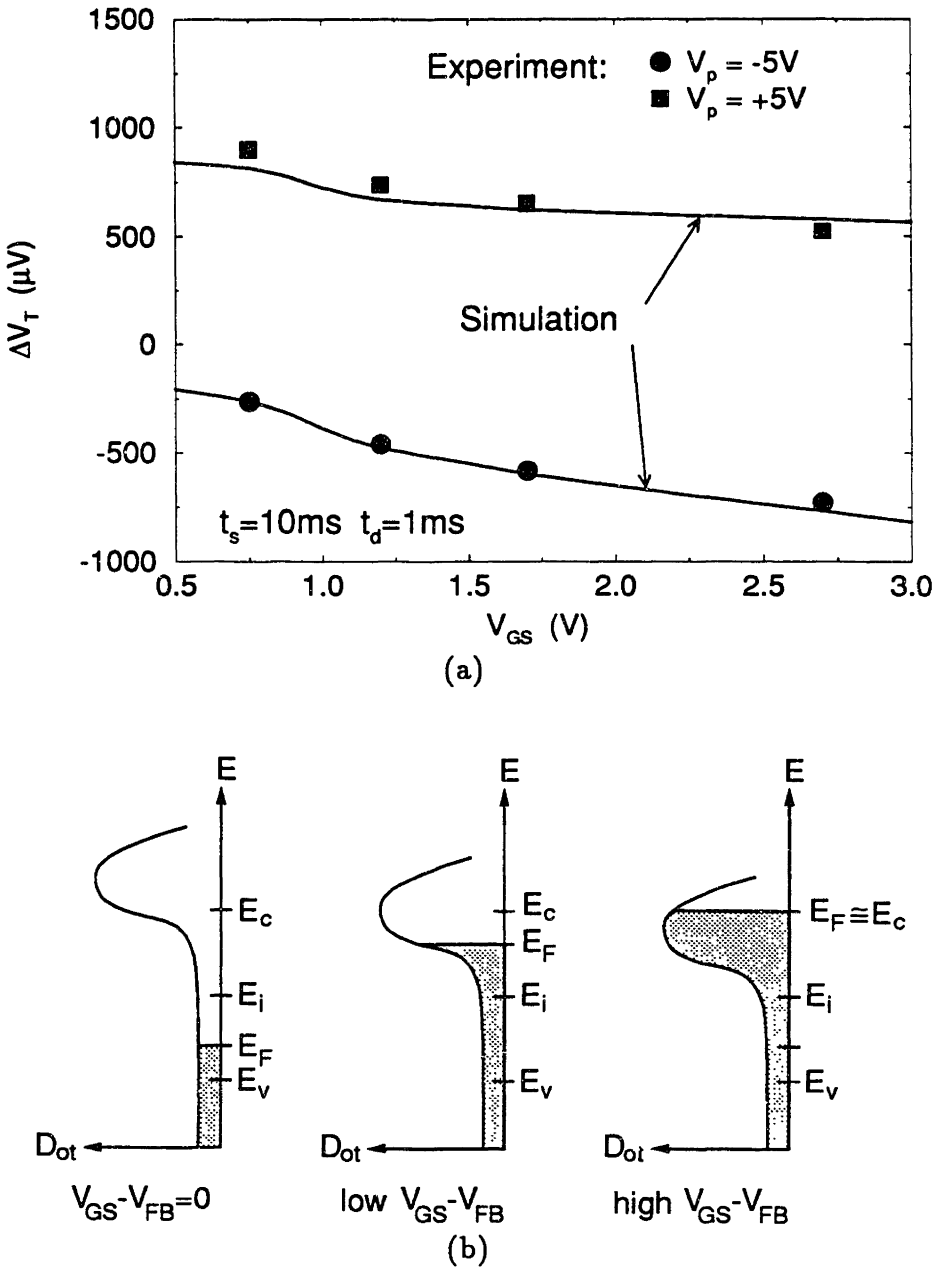


Figure 7.45: (a) Initial gate bias dependence; simulated and measured. (b) Change in trap occupancies as a function of gate bias.

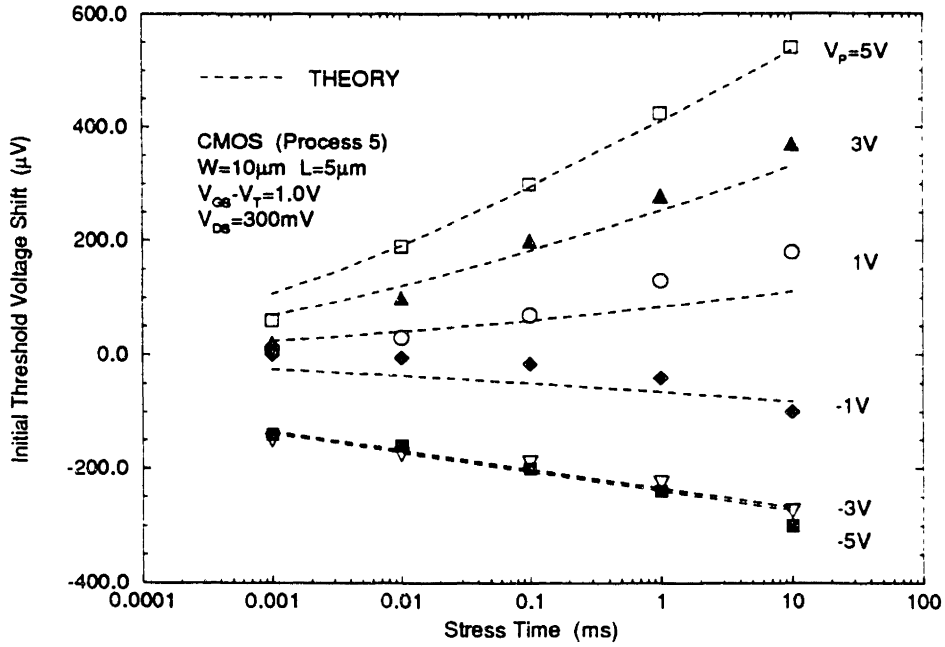


Figure 7.46: Simulated and measured V_T shifts versus stress time with stress voltage V_p as a parameter.

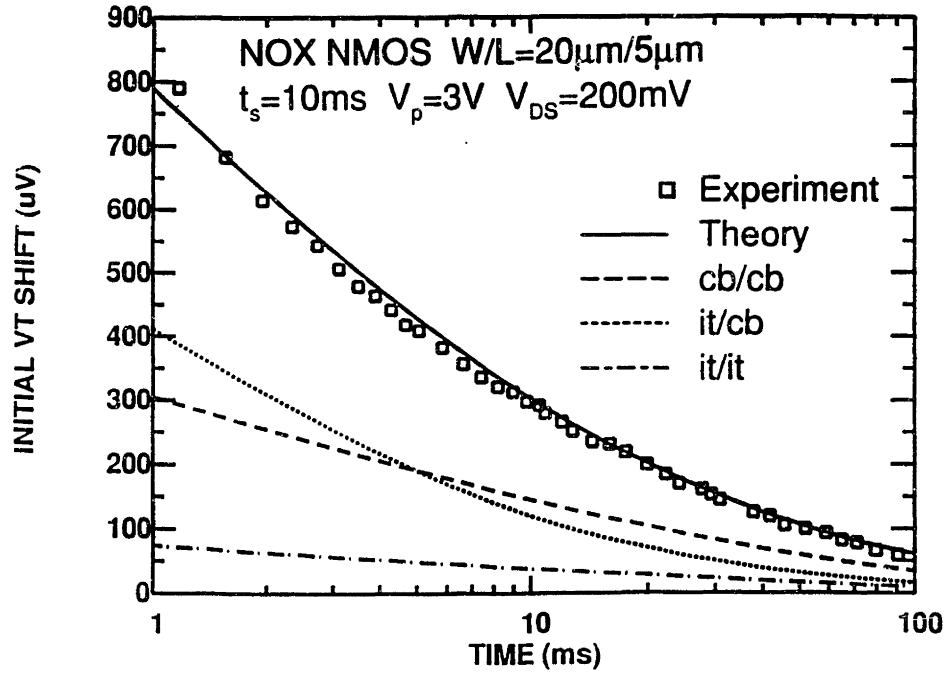
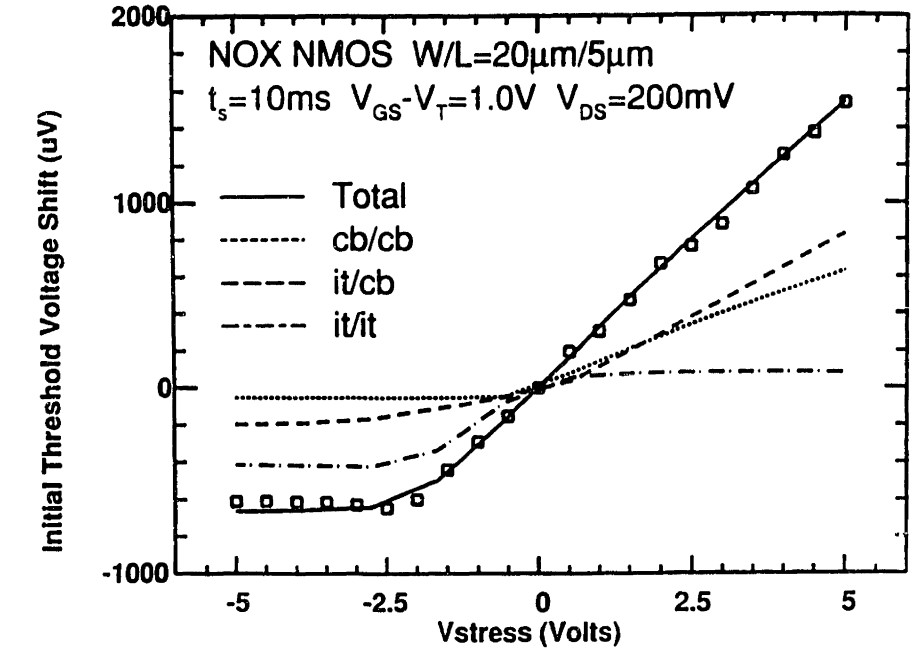
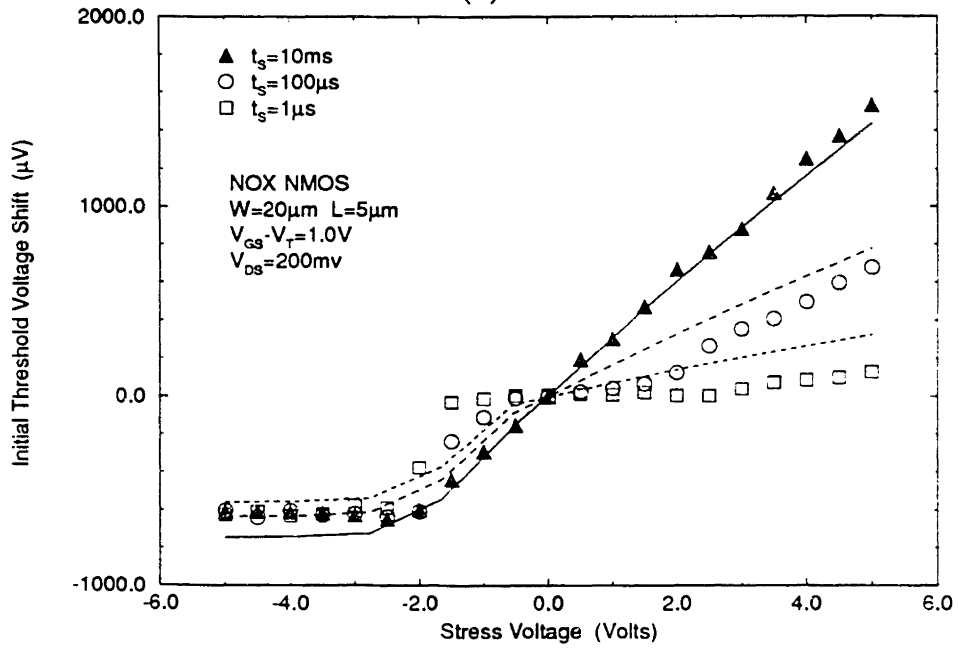


Figure 7.47: Simulated relaxation transient showing decomposition by mechanism.



(a)



(b)

Figure 7.48: Dependence of initial threshold voltage shift on pulse amplitude, NOX NMOS 20/5, $V_{GS} - V_T = 1.1V$, $V_{DS} = 200\text{mV}$ (a) Breakdown by component (b) Parameterized by stress time.

in nitrided devices were larger than for conventional SiO_2 dielectrics. The trap densities D_1 and D_0 in Table 7.6 show good correlation with observed threshold voltage shifts as given in Tables 7.2 and 7.4 under positive and negative gate voltage pulses, respectively. They also show good agreement with oxide trap distributions extracted using 1/f noise measurements [51, 214].

We make one last observation on the apparent reduction in the oxide trap density with energy below the conduction band edge. This observed decrease may not in fact be real, but may appear artificially low as a result of the screening or "filtering" action of the interface states through which the oxide traps charge in the two-step mechanism. Above the conduction band, where inversion layer electrons flood the interface, all of the oxide traps are accessible to tunneling electrons, whereas below the conduction band, only a small fraction of the traps are accessible, as determined by the probability that an interface state and an oxide trap happen to fall within tunneling distance. Since this probability is not taken into account in the model, the oxide trap density below the conduction band may be artificially low.

7.2.5 Relation to 1/f Noise

We have postulated that threshold voltage relaxation is a large-signal manifestation of the tunneling of channel charge carriers to and from near-interface oxide traps. Since 1/f noise is a small-signal manifestation of the same mechanism, it follows that there should be a strong correlation between the two. Such a correlation has already been demonstrated in a previous study [10]. In this section we pursue this relation further.

In Chap. 4 expressions were derived for the various components of the trapping-induced threshold voltage shift. In the tunneling-limited regime where $\tau_{ot/it} \gg \tau_{it}$, e.g. in strong inversion, all of these components reduce to simple first order rate equations and can be lumped into one expression of the form

$$\Delta V_T \approx \frac{q}{C_{ox}} \int_0^{t_{ox}} dx \int dE D_{ot}(E,x)[f(E+q\mathcal{E}x) - f_0(E+q\mathcal{E}_o x)] \times \left[1 - e^{-t_r/\tau_{ot}(x)}\right] e^{-t_r/\tau_{ot}^o(x)} \tag{7.5}$$

where it has also been assumed that the trapped charge is close enough to the interface so that $(t_{ox} - x)/t_{ox} \approx 1$. We shall be interested in strong inversion, where the tunneling occurs near the silicon conduction band and the oxide trap time constant is given by Eq. 4.51,

$$\begin{aligned} \tau_{ot}(x) &= \tau_o \exp(x/x_o) \\ x_o &= \frac{\hbar}{2\sqrt{2M^*\phi_c}} \end{aligned} \tag{7.6}$$

For the Si/SiO_2 system with $M^* = 0.42m_o$, and $\phi_c = 3.13$eV, the characteristic tunneling depth is $x_o = 0.85\mathring{A}$.

In a model of 1/f noise based on carrier number fluctuations, the power spectral density of the equivalent gate voltage noise is [29]

$$\overline{V_{GS}{}^2} = \frac{q^2}{WLC_{ox}{}^2}\int_0^{t_{ox}} dx \int_{E_v}^{E_c} dE\, D_{ot}(E,x) f(E)[1-f(E)] \frac{\tau_{ot}(x)}{1+\omega^2\tau_{ot}{}^2(x)} \tag{7.7}$$

with τ_{ot} given by Eq. 7.6. The factor $f(E)[1-f(E)]$ resembles an impulse centered around $E = E_{Fn}$ so that the integral in Eq. 7.7 can be carried out, giving

$$\overline{V_{GS}{}^2} = \frac{q^2 x_o}{WLC_{ox}{}^2}\frac{kTD_{ot}(E_{Fn})}{f} \tag{7.8}$$

Thus, the 1/f noise probes the oxide trap density immediately adjacent to the electron quasi-Fermi level in the silicon. In contrast, the probe function $\Delta f = f - f_0$ in Eq. 7.5 resembles a broadened pulse as shown in Fig. 4.16. Thus V_T relaxation contains information on the oxide trap density over a much wider range of energies than does 1/f noise. Nevertheless, since the expressions 7.5 and 7.7 for the threshold voltage shift and the mean squared gate voltage noise are both directly proportional to the oxide trap density, they are expected to be closely correlated as long as $D_{ot}(E,x)$ varies relatively smoothly with energy over the range of interest. However, if $D_{ot}(E,x)$ varies rapidly with energy, then the V_T relaxation measurement may probe regions of greater or lesser trap densities than are accessible to the 1/f noise measurement, weakening or destroying the correlation.

Figure 7.49 supports this theory. Threshold voltage relaxation transients are shown in Fig. 7.49(a) for 20μm/5μm NMOS transistors in three different processes. The devices are biased initially with $V_{GS} - V_T$=0.5V and V_{DS}=0.5V and pulsed with V_p=5V for a period of 10ms. Figure 7.49(b) shows 1/f noise measured in these same devices under identical bias conditions. The 1/f noise measurements were obtained using a Quan-Tech 5173 Semiconductor Noise Analyzer and an HP 3561A Dynamic Signal Analyzer. A strong qualitative correlation between 1/f noise and the trapping-induced V_T shift is readily apparent.

In order to quantitatively correlate the two sets of measurements, we must take into account the geometry dependence of the constant of proportionality. For a given set of stress and bias conditions, Eqs. 7.5 and 7.7 imply

$$\Delta V_T \propto \frac{WLC_{ox}}{q}\overline{V_{GS}{}^2} \tag{7.9}$$

This does not take into account the fact that, for a given V_p, the oxide band bending and hence ΔV_T depend on t_{ox}. We shall correct for this below. In order to get some idea of the correlation, Fig. 7.50 plots $\frac{WLC_{ox}}{q}\overline{V_{GS}{}^2}$ (V) at 100 Hz together with the initial V_T shifts (t_d=1ms) for $V_p = \pm 5$V in five different processes. Values are also given for the correlation coefficients r_+ and r_- of the 1/f noise data with the positive

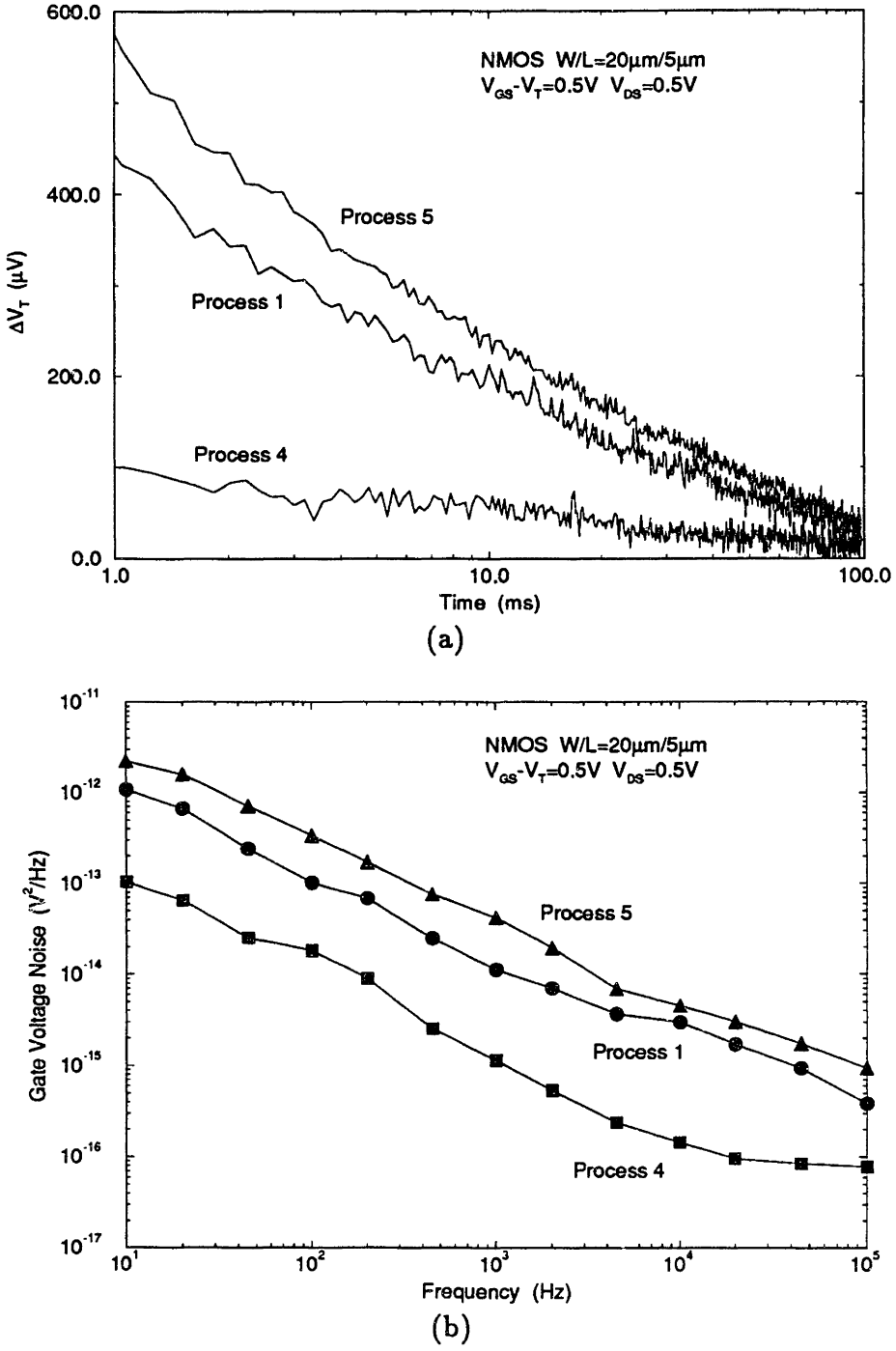


Figure 7.49: Comparison of three different 20/5 NMOS transistors (a) threshold voltage relaxation transient (b) 1/f noise.

and negative threshold voltage shift data respectively. The correlation coefficient between x and y is defined by [215]

$$r = \frac{E(xy) - E(x)E(y)}{\sigma_x \sigma_y} \tag{7.10}$$

where $E(x)$ is expected value of x and σ_x its standard deviation. A high degree of correlation is evident in both NMOS and PMOS devices. PMOS transistors exhibit very much lower levels of both 1/f noise and V_T shifts [note the scale change in Fig. 7.50(b)]. Deviations may be explained by nonuniformities in the trap distribution, as well as by experimental errors in the determination of the small V_T shifts, especially for PMOS devices. Note that the correlation is stronger between 1/f noise and the V_T shift in the direction of strong inversion (positive pulses for NMOS, negative pulses for PMOS). This is reasonable since, in strong inversion, the range of trap energies probed during the 1/f and V_T measurements are more nearly equal due to the pinning of the Fermi level, i.e. the threshold voltage probe $\Delta f = f - f_o$ reduces to the 1/f noise probe $f_o(1 - f_o)$.

For pulses of the opposite polarity, the Fermi level sweeps across the bandgap and the threshold voltage measurement probes traps far outside the range which contributes to 1/f noise.

In order to obtain a first-order estimate of the actual value of the threshold voltage shift expected from a device with a given 1/f noise, we begin by assuming a uniform trap distribution $D_{ot}(E, x) = D_{ot}$. As shown in Sec. 4.4, Eq. 7.5 can then be written

$$\begin{aligned} \Delta V_T &= \frac{qD_{ot}}{C_{ox}}(E_F - E_{F0}) \int_0^{t_{ox}} \left[1 - e^{t_s/\tau_{ot}(x)}\right] e^{-t_r/\tau_{ot}(x)} dx \\ &\quad - \frac{q(\mathcal{E} - \mathcal{E}_o)D_{ot}}{C_{ox}} \int_0^{t_{ox}} x \left[1 - e^{t_s/\tau_{ot}(x)}\right] e^{-t_r/\tau_{ot}(x)} dx \end{aligned}$$

These are just the it/it and it/cb components of Sec. 4.4. The cb/cb component does not enter because we assumed a step Fermi function. In strong inversion $E_F \approx E_{F0}$ so that the first (it/it) component does not contribute and the second integration can be carried out to give

$$\Delta V_T \approx \frac{q^2 D_{ot} {x_o}^2}{2\epsilon_{ox}} V_p \left[\ln^2\left(\frac{t_s}{\tau_o}\right) - \ln^2\left(\frac{t_r}{\tau_o}\right)\right] \tag{7.11}$$

From Eqs. 7.8 and 7.11, the relation between 1/f noise and threshold voltage shift in strong inversion can be written

$$\Delta V_T = C\overline{V_{GS}{}^2} f V_p \left[\ln^2\left(\frac{t_s}{\tau_o}\right) - \ln^2\left(\frac{t_r}{\tau_o}\right)\right] \tag{7.12}$$

where C (V^{-2}) is a constant given by

$$C = \frac{x_o \epsilon_{ox} WL}{2kTt_{ox}^2} \tag{7.13}$$

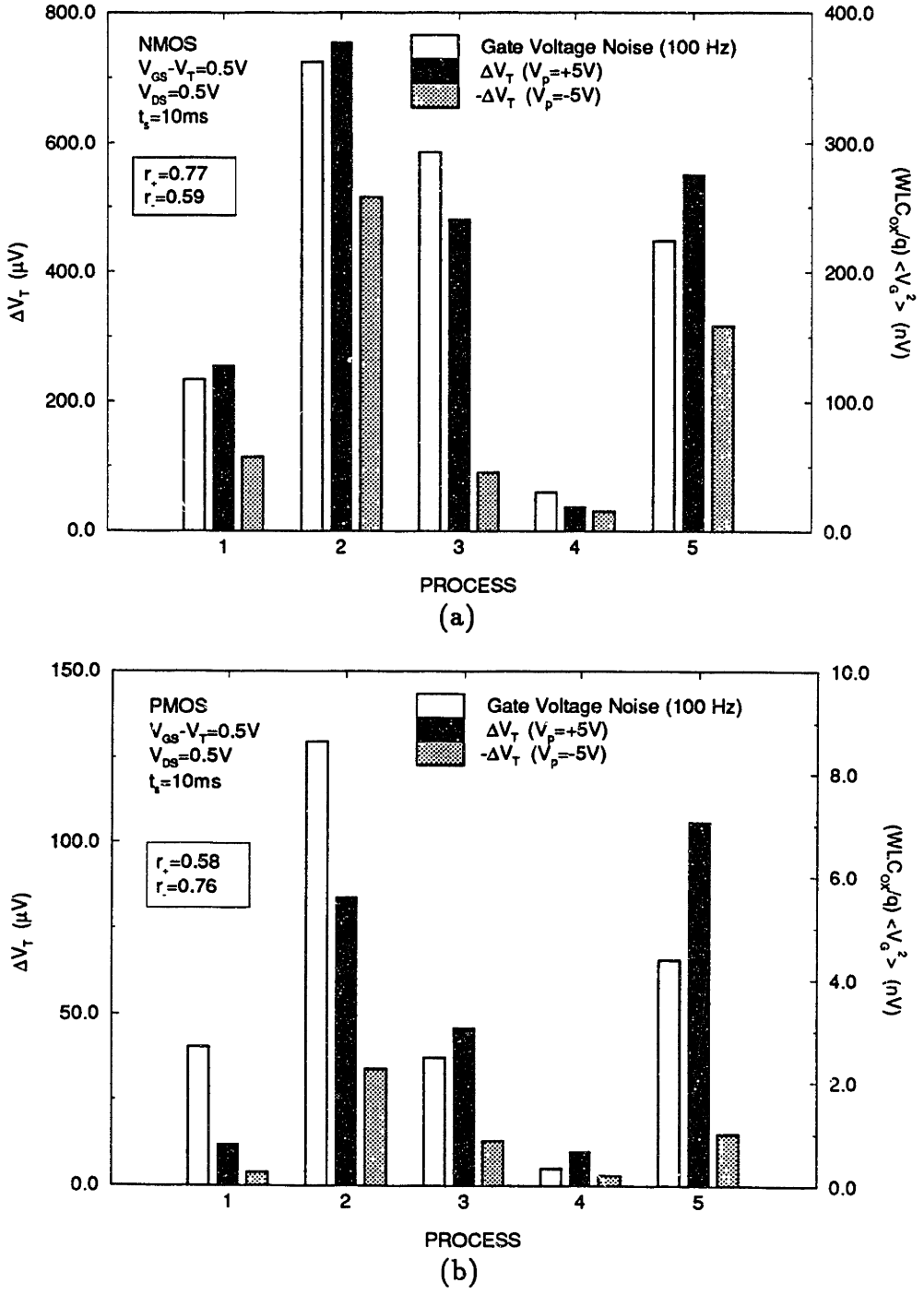


Figure 7.50: 1/f noise at 100 Hz and peak V_T shifts for various processes (a) NMOS (b) PMOS.

Table 7.7: Comparison of measured V_T shifts with estimates based on 1/f noise.

		NMOS			PMOS		
			ΔV_T ($V_p = +5$V)			ΔV_T ($V_p = -5$V)	
Process	C	$\overline{V_{GS}^2}^{1/2}$	measured	predicted	$\overline{V_{GS}^2}^{1/2}$	measured	predicted
1	3.0e4	436	255	227	66	-4	-5
2	1.5e5	343	754	701	53	-34	-13
3	1.25e4	728	481	263	67	-13	-2
4	1.47e4	263	38	40	28	-3	-0.5
5	5.87e4	510	551	607	71	-14	-12

Using $\tau_o = 10^{-10}$s [29] and $x_o = 0.85\mathring{A}$, the threshold voltage shift for V_p =5V at t_s =10ms and t_d =1ms can be estimated from the gate voltage noise at one frequency. Table 7.7 lists the gate voltage noise at 100 Hz, the value of C, and the measured and estimated V_T shifts. Since the approximation is only valid in strong inversion, V_T shifts for NMOS and PMOS transistors are tabulated only for $V_p = +5$V and -5V, respectively. The estimated and measured V_T shifts agree well within the approximations of the analysis and experimental error.

A strong correlation between 1/f noise and threshold voltage relaxation has been demonstrated. However, in the above discussion we have glossed over differences in the various components of the V_T shifts as well as nonuniformities in the oxide trap spatial and energy distributions. These factors can be rigorously accounted for through a comparison of the trap spatial and energy distributions extracted using each of the two techniques. In principle, V_T relaxations can be measured over a range of biases to reconstruct the oxide trap distributions in both energy and space. Jayaraman et al [214, 51] have devised a 1/f noise technique to measure the oxide trap density near the silicon conduction band edge. This technique has been used to measure the density of traps at a depth of $5 - 9\mathring{A}$ from the interface over an energy range $25 < E - E_C < 40$meV. Reported trap densities in NMOS devices are on the order of 1×10^{17}cm^{-3} eV^{-1} in conventional oxides and $5 \times 10^{17} - 1 \times 10^{18}$cm^{-3} eV^{-1} in NOX and ROXNOX dielectrics [51]. These values are consistent with the trap densities required to fit threshold voltage relaxation data; however further research in this area is needed.

7.2.6 Circuit Issues

Model

The effects of threshold voltage relaxation on circuit performance depend on circuit topology, the speed and accuracy required, as well as on bias conditions and the specific sequence of inputs to which the device of interest is subjected. In order to

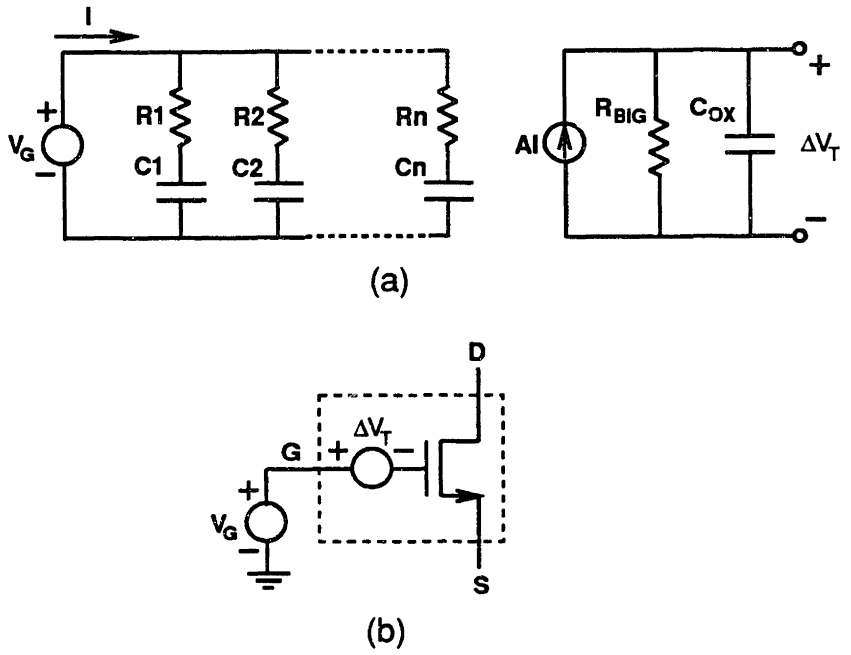


Figure 7.51: Simple SPICE subcircuit model for V_T relaxation.

obtain a quantitative estimate of the impact of trapping-induced threshold voltage shifts in a specific application, it is desirable to have a simple model for use in a circuit simulation environment such as SPICE. Such a model need not be based on rigorously correct physics, but must be computationally efficient and characterized by relatively few parameters which can be easily extracted from a small set of measured data.

Figure 7.51 shows a simple circuit model with which threshold voltage relaxation can be simulated. The topology is based on Dow's model for dielectric relaxation in a capacitor, as discussed in Sec. 5.2. Each of the N rungs of the RC ladder represents an oxide trap, with increasing time constants corresponding to deeper traps according to Eq. 4.51. In response to a gate voltage pulse $\Delta V_G = V_p$, the current which charges the capacitors is given by

$$I(t) = \sum_{j=1}^{N} \frac{V_p}{R_j} e^{-t/\tau_j} \tag{7.14}$$

where $\tau_j = R_j C_j$. The resulting threshold voltage shift, assuming all the trapped charge Q_{ot} is close to the interface, is

$$\Delta V_T(t) = -\frac{Q_{ot}(t)}{C_{ox}} = \frac{1}{C_{ox}} \int_0^t I(t)dt \tag{7.15}$$

The integration is performed by the second network consisting of a dependent current source with gain A, charging a capacitor C_{ox} shunted by a large resistor R_{BIG} to

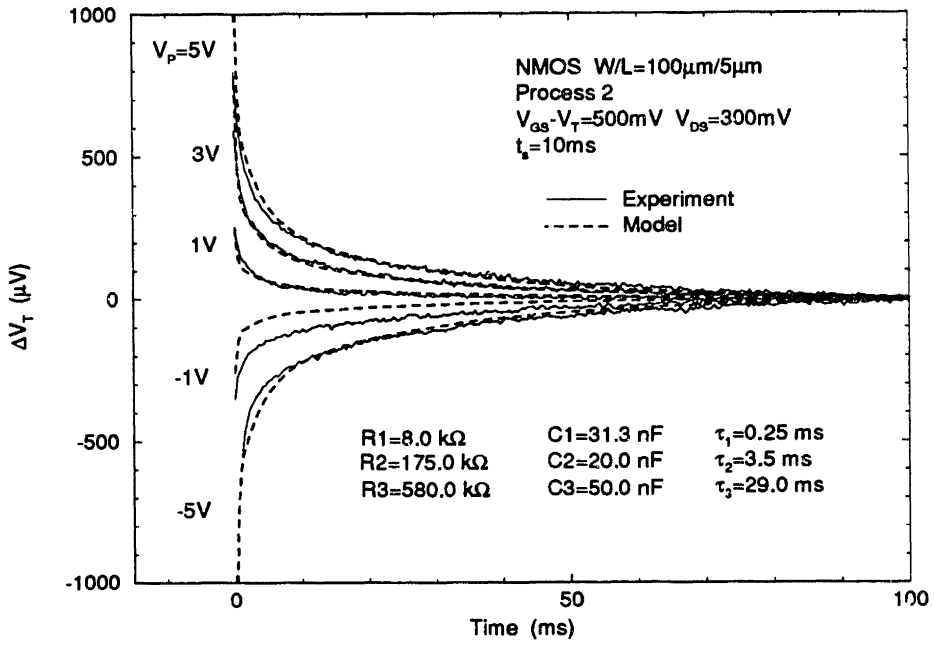


Figure 7.52: Measured V_T relaxation data and simulations using SPICE model (t_s = 10ms).

provide a DC path to ground. The output is

$$\Delta V_T = \frac{AV_p}{C_{ox}} \sum_{j=1}^{N} C_j \left[1 - e^{-t/\tau_j}\right] \tag{7.16}$$

A dependent voltage source in series with the gate terminal of the MOSFET reduces the external V_{GS} by the amount ΔV_T as shown in Fig. 7.51(b), effectively increasing the threshold voltage.

It was shown in Sec. 4.4 that in strong inversion the threshold voltage shift is given by

$$\begin{aligned} \Delta V_T &= \frac{q^2 D_{ot}}{\epsilon_{ox}} V_p \int_0^{t_{ox}} x \left[1 - e^{-t/\tau(x)}\right] dx \\ &\approx \frac{q^2 D_{ot}}{\epsilon_{ox}} V_p \sum_{j=1}^{N} \left[1 - e^{-t/\tau(x)}\right] x_j \Delta x_j \end{aligned} \tag{7.17}$$

The constants A and component values C_j and R_j can be chosen to make Eqs. 7.16 and 7.17 equal for a given partioning of the oxide into slices Δx_j. In practice these parameters are best determined by iteration to achieve a required degree of fit to a set of experimental data.

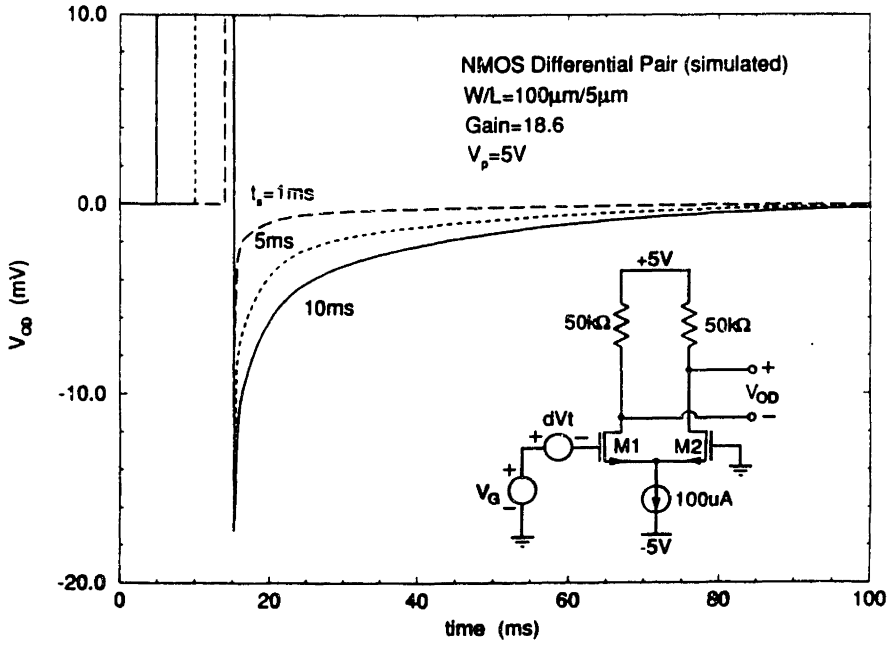


Figure 7.53: Simulated pulse response of a differential pair using the SPICE V_T relaxation model.

Figure 7.52 shows a set of V_T relaxation transients at several values of V_p together with simulations using the model of Fig. 7.51. Three time constants, spaced about a factor of ten apart, were sufficient to fit the data over the time range shown. In order to accurately model shorter time scales, it may be necessary to add more rungs to the ladder with shorter time constants. The agreement of model with experiment for positive V_p is quite good. Since the model is symmetrical however, it fails to model the nonlinearity and saturation for negative voltage pulses. Nevertheless, the quality of the fit is probably satisfactory for most circuit simulation purposes.

Improvements to the model could be easily made by making the dependent V_G source in Fig. 7.51(a) a nonlinear function of V_G in the gate terminal [Fig. 7.51(b)]. For example, an ideal diode could be used to implement the unipolar ΔV_T–V_p characteristic of a device with a backgate bias.

Figure 7.53 shows an example of the use of the model to simulate the output of an NMOS differential pair subjected to gate voltage pulses of amplitude 5V of varying durations. During the pulse, the differential output voltage V_{OD} goes positive. When the pulse is removed at time t = 15ms, the output does not return to zero, but overshoots in the negative direction due to electrons trapped in the oxide of M1, increasing its V_T. The amplitude of the initial overshoot increases with stress time and pulse amplitude, and decays with a long tail, as observed in measurements of MOS differential pairs [10, 62].

While the V_T relaxation model increases circuit simulation time somewhat, it is

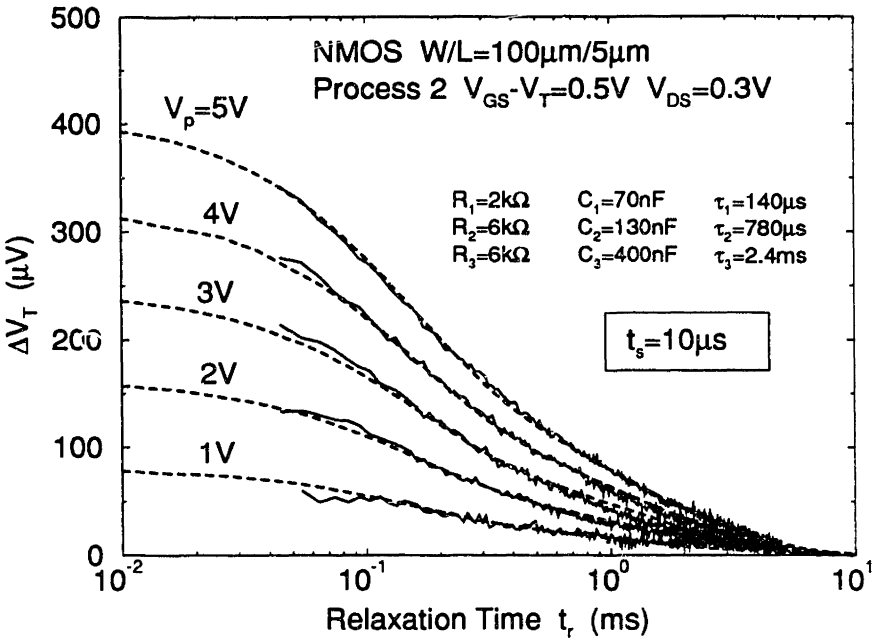


Figure 7.54: Measured and simulated relaxation transients in process 2 using the SPICE model ($t_s = 10\mu s$).

only necessary to include the model in critical devices. For example, in a comparator or operational amplifier, it is normally sufficient to model only the input devices since V_T shifts in later stages are usually negligible when referred to the input.

Effect on Circuit Accuracy

As discussed above, the effects of threshold voltage relaxation on circuit performance depend on the specifics of the circuit architecture and input waveform. The best way to evaluate the effects of threshold voltage relaxation is by simulation under a range of worst-case inputs, using a model such as that presented above. In order to get some idea of the order of magnitude of the error expected, consider an analog-to-digital converter with full-scale voltage V_{REF} operating at a clock frequency f_{CLK}. As a figure of merit, we take the threshold voltage shift after a delay $t_r = 1/f_{CLK}$, following stress at a voltage $V_p = V_{REF}/2$ for a duration $t_s = 1/f_{CLK}$. Thus, higher clock frequencies allow less time for traps to fill during stress, but also allow less time for emission before sampling the threshold voltage shift.

As an example, in most analog-to-digital converters, the input stage of a comparator will be subjected to half the reference voltage for one clock period during the MSB comparison. If the input device has not recovered from this stress by the end of the following clock period, i.e. if there is a residual threshold voltage shift, an error

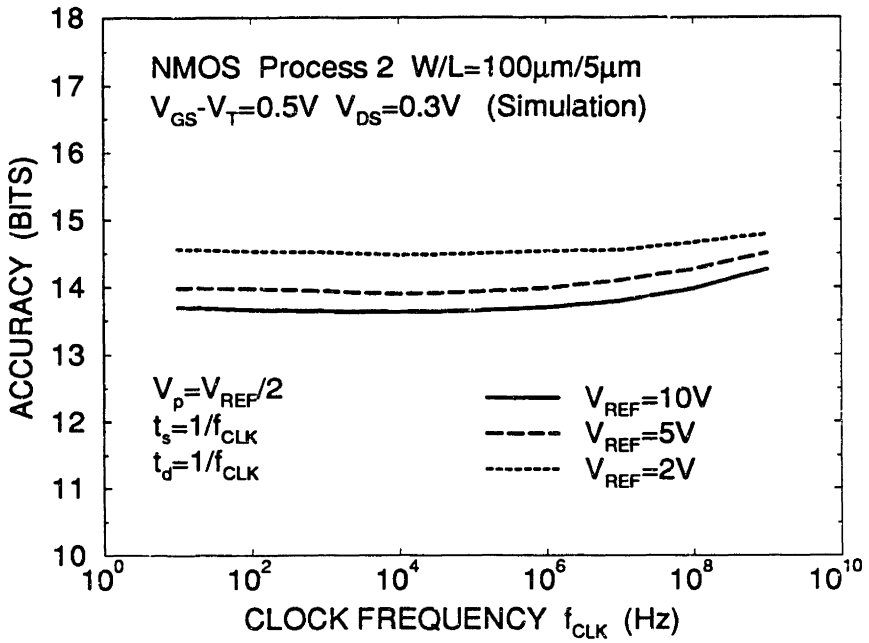


Figure 7.55: Accuracy (in bits) versus clock frequency with reference voltage as a parameter.

can occur on the next bit decision. Other types of circuits may be subjected to more severe stress conditions, as for example a sample-and-hold, in which the magnitude and duration of the stress depends on the specific input being sampled. Our figure of merit does not represent a worst case stress condition but provides a reasonable basis for estimating circuit performance.

The closed-loop measurement circuit cannot recover from stress fast enough to accurately observe the threshold voltage shift 1-10μs after stress, but simulations enable us to do so. Figure 7.54 shows the model of Fig. 7.51 fitted to data from an NMOS transistor in Process 2 with the parameters shown. From the figure, it is clear that the threshold voltage shift 10μs after 5V of stress for 10μs is about 390μV. Since an LSB is $V_{REF}/2^N$, for $\pm 1/2$ LSB relative error, this represents an accuracy of 13.65 bits.

Figure 7.55 plots accuracy, in number of bits at the $\pm 1/2$ LSB level, as a function of clock frequency using

$$N = 3.322 \log_{10} \left(\frac{V_{REF}}{\Delta V_T (1/f_{CLK})} \right) - 1$$

where N is the number of bits. The figure was generated with the exact numerical simulation model using the parameters in column 3 of Table 7.6. The results show that, for this process, the error due to threshold voltage relaxation is on the 13-

14 bit level. The device chosen represents a "typical" sample and results will vary considerably among processes, as shown in Table 7.2.

The accuracy increases slightly with decreasing reference voltage, indicating that the V_T shift falls off faster with decreasing stress voltage than does the size of the LSB. The reference voltage dependence is not strong however, since the threshold voltage shift is approximately proportional to stress voltage as shown in Fig. 7.54 and discussed in Sec. 4.4. Nonuniformities in the trap distribution may lead to additional reference voltage dependence.

The accuracy is relatively insensitive to frequency over most of the range of interest (10kHz-1MHz), due to the tradeoff between increasing stress time, which increases the time available for traps to charge and increasing relaxation time, which allows more emission to take place. The simulation indicates that accuracy increases at very high frequencies when the traps do not have time to respond to any significant degree.

Minimization of the Threshold Voltage Shifts

The most direct way to minimize threshold voltage relaxation effects in circuits is to eliminate the source of the problem by reducing the density of oxide traps and interface states through improved processing technologies. As indicated in Tables 7.2–7.5 the variation of the magnitude of the threshold voltage relaxation between different processes may span one or two orders of magnitude. Further work is required in order to correlate these differences with processing conditions. In any case, threshold voltage relaxation effects should be an important consideration in choosing a process for sensitive analog applications. Since this degree of freedom is usually not available to the circuit designer, we seek bias conditions for the minimization of the threshold voltage shifts.

It has been shown that the application of a substrate bias greatly reduces the magnitude of the threshold voltage shift for pulses which drive the device into accumulation. Strong inversion operation can, to a lesser extent, suppress the threshold shift in the direction of inversion. The maximum benefit that can be obtained through adjustment of bias conditions occurs when these conditions are satisfied simultaneously. Figure 7.56(a) shows the relaxation transient for an NMOS transistor stressed with $\pm$5V for 10μs. The application of a substrate bias eliminates most of the negative V_T shift while operation in very strong inversion reduces the positive shift by about a factor of two.

In most conventional SiO_2 dielectrics, PMOS devices exhibit lower threshold voltage shifts than NMOS, so that PMOS devices should be used whenever possible in critical areas such as input stages of comparators or operational amplifiers. There are exceptions to this rule, however, as exemplified by NOX and ROXNOX dielectrics.

When the above bias conditions are used in conjunction with PMOS devices, the minimum threshold voltage shifts are obtained. Figure 7.56(b) shows a PMOS device stressed with $\pm$5V for 10ms. This device is typical of most of the PMOS transistors

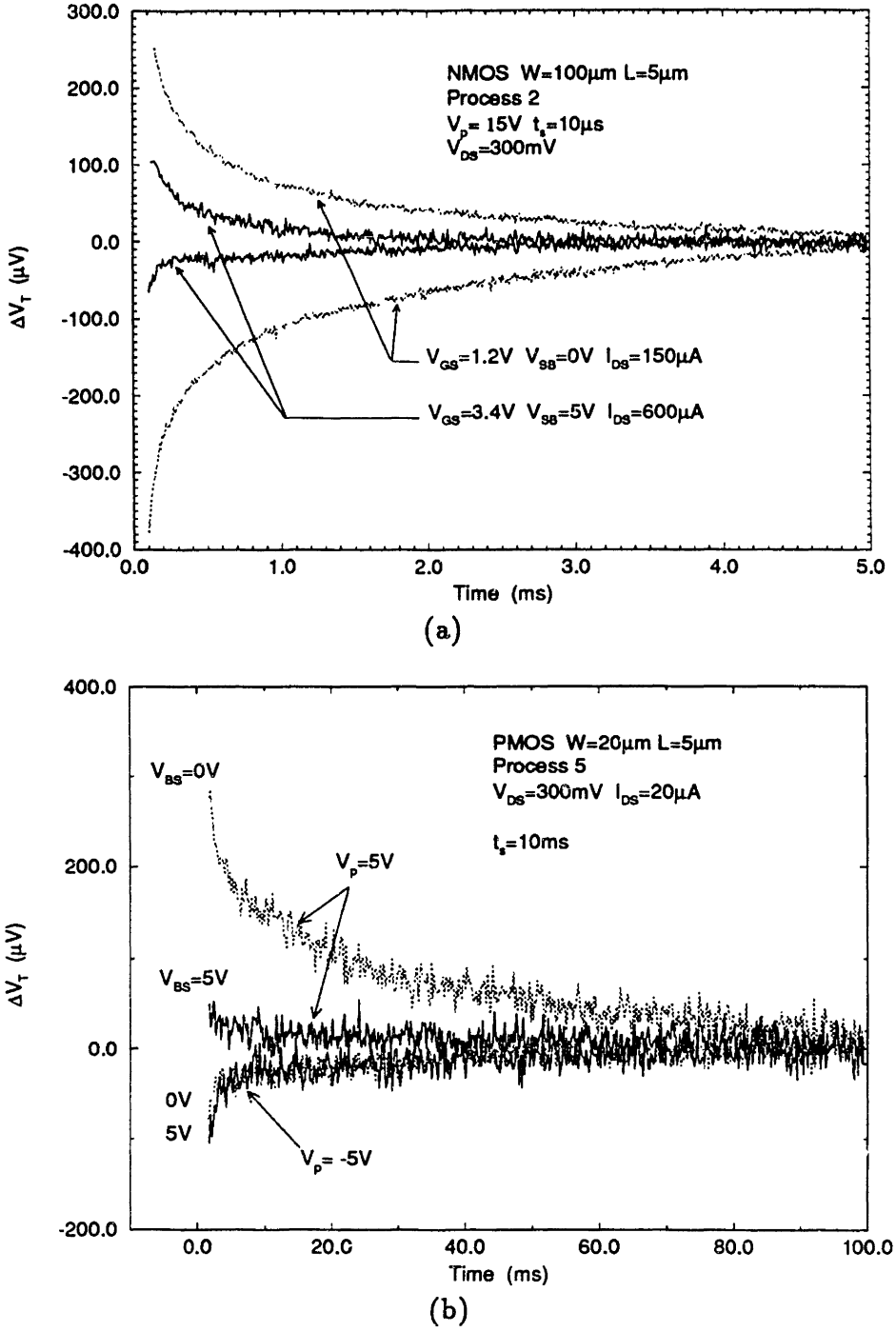


Figure 7.56: Suppression of V_T shift by operating in strong inversion with a backgate bias (a) NMOS (b) PMOS.

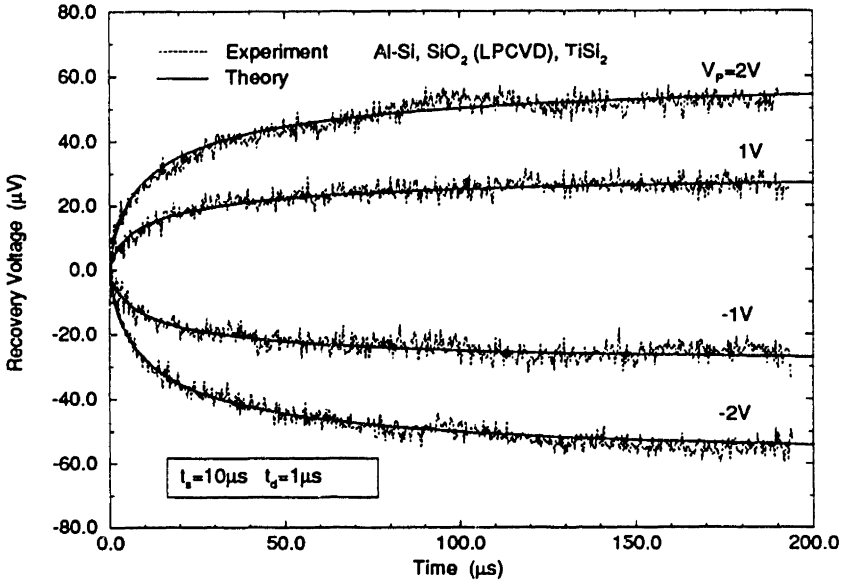


Figure 7.57: Recovery voltage vs. relaxation time for a capacitor with a LPCVD SiO_2.

shown in Table 7.2 in that the threshold voltage shift is greater for positive than for negative stress voltages, consistent with electron trapping. This situation is very fortunate, because the simple application of a backgate bias then results in a device with minimal V_T shifts in both directions, as shown in Fig. 7.56(b). This fix is not guaranteed to work, since some devices may exhibit larger shifts in the negative direction, e.g. process 3 in Table 7.2, however it has been found to be effective in most of the devices tested. It is important, when using this technique, that the backgate bias be comparable to the largest anticipated stress voltage so as not to cross over the step in ΔV_T–V_p curve.

7.3 Dielectric Relaxation in Capacitors

In this section we present measurements of dielectric relaxation in capacitors together with simulated results using the models developed in Chap. 5. Two types of dielectric were tested. One set of capacitors has an LPCVD SiO_2 dielectric between a $TiSi_2$ bottom plate and a Al-Si top plate. The second structure consists of a $TiSi_2$ bottom plate followed by 600$\AA$ LPCVD SiO_2, 150$\AA$ Si_3N_4 and a Ti-W top plate. The devices for this study were provided by Texas Instruments Inc. and are described in greater detail by Fattaruso et al [166]. In SiO_2 capacitors stressed for 10μs and discharged for 1μs, the recovery voltage that develops as a function of time after the capacitor

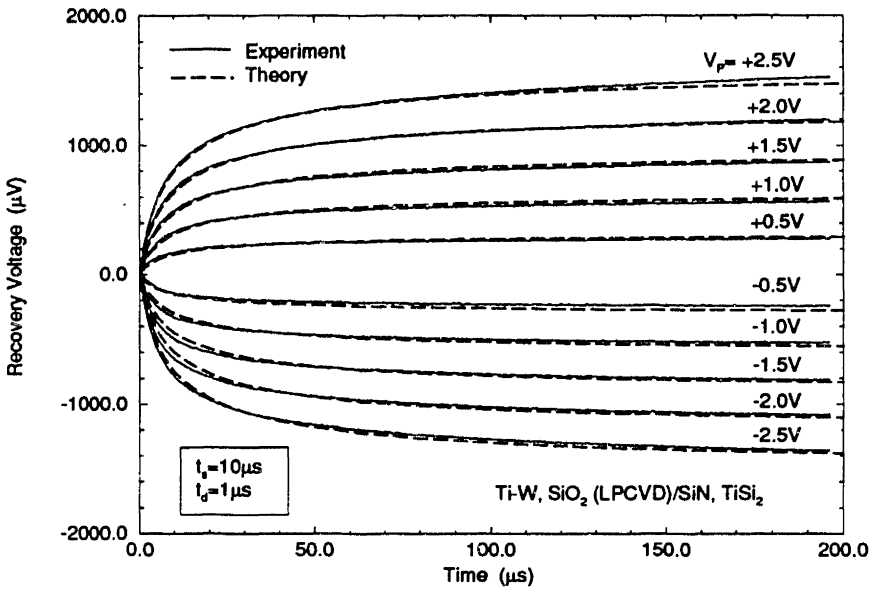


Figure 7.58: Recovery voltage vs. relaxation time for a capacitor with a nitrided oxide dielectric.

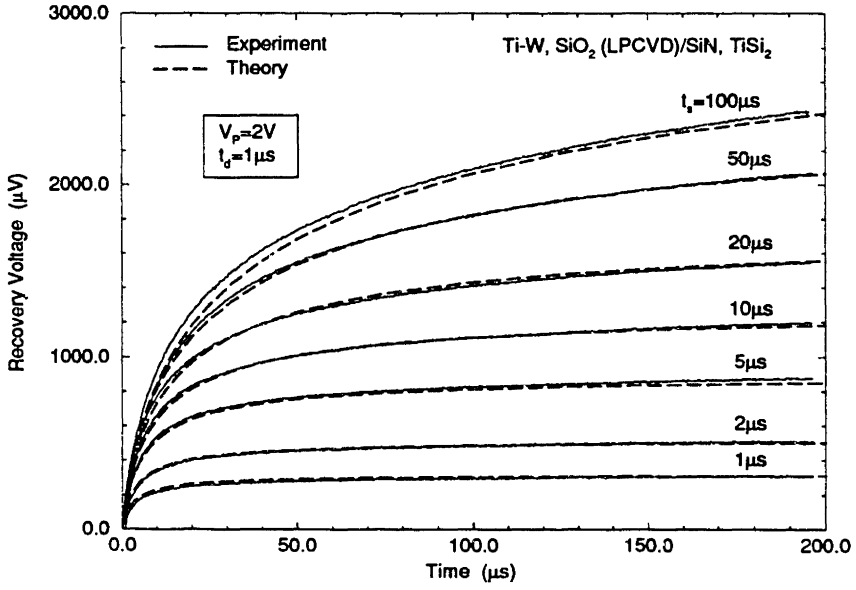


Figure 7.59: Recovery voltage vs. relaxation time with stress time (t_s) as a parameter.

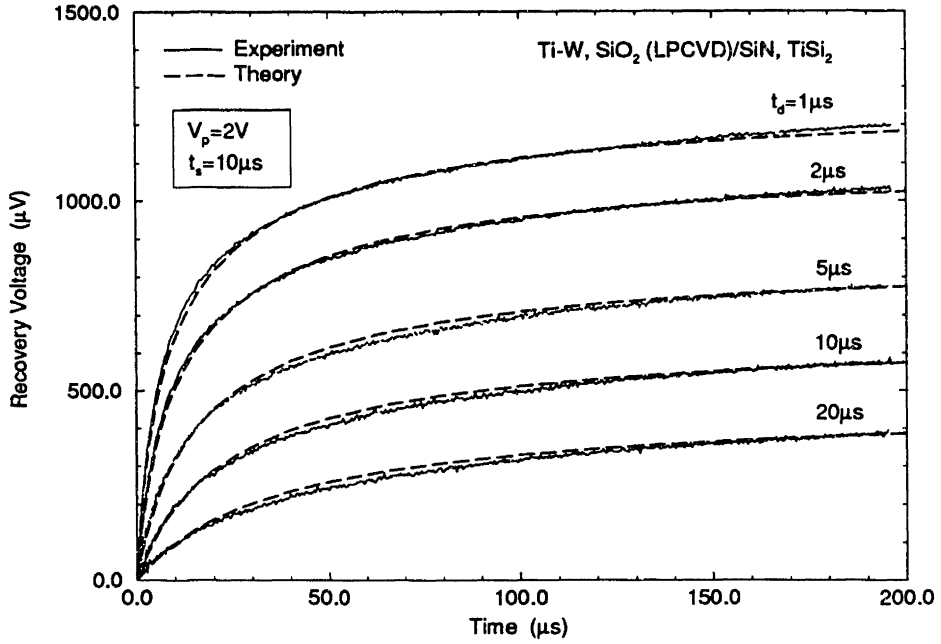


Figure 7.60: Recovery voltage vs. relaxation time with discharge time (t_d) as a parameter.

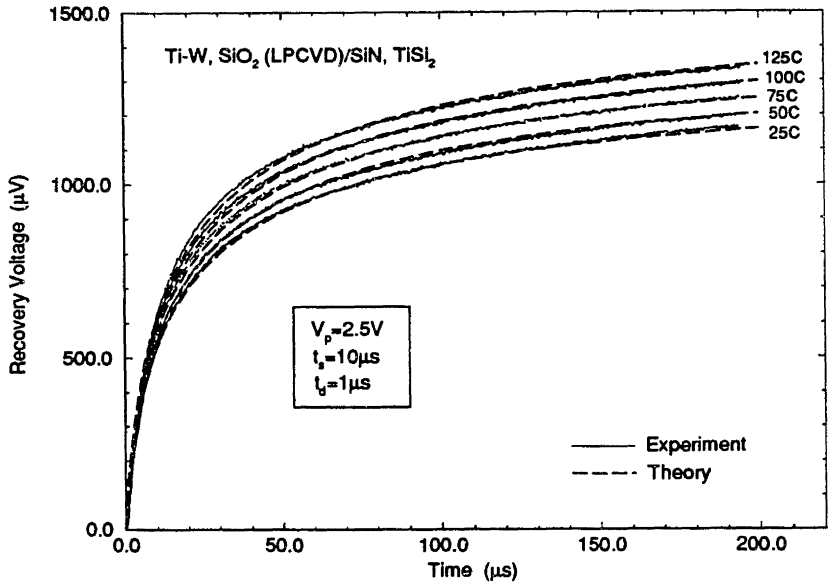
is open-circuited is shown in Fig. 7.57. Also shown is the simulated recovery voltage using Eq. 5.30 with

$$\begin{aligned} D_{ot} &= 1.7 \times 10^{17}\text{cm}^{-3}\ \text{eV}^{-1} \\ \tau_{oc} &= 1.0\text{ps} \\ \tau_{oe} &= 14.3\text{ps} \\ x_o &= 0.74\mathring{A} \end{aligned}$$

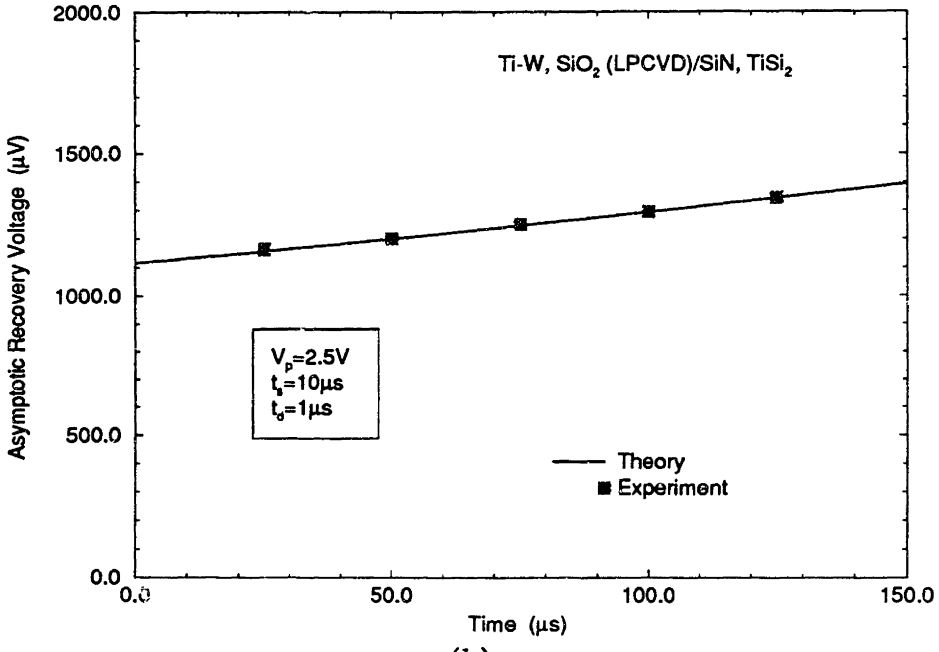
We have seen in Fig. B.4 that the actual value of the time constant depends on the energy of the tunneling electron. Nevertheless, we shall follow the traditional practice of using an "average" value to serve as a fitting parameter in the model, keeping in mind that by doing so the parameter loses to some extent its physical meaning. Typical values of τ_o reported in the literature range from 10^{-13}–10^{-8} s [73, 72, 216, 29] for SiO_2 and may be as high as $5 \times 10^{-4}s$ for polymers [197].

The data shows a good fit to the linear voltage dependence and the $\log^{-3}$ time dependence predicted by the model. The asymptotic recovery voltage, defined as the final value of the recovery voltage, expressed as a fraction of the stress voltage, is approximately 27ppm. For an A/D converter with a full-scale voltage of 5V, this represents an error at the 16-bit level.

Experimental (solid line) and simulated (dotted line) results for nitrided dielectrics under the same stress conditions are shown in Fig. 7.58. The asymptotic recovery



(a)



(b)

Figure 7.61: Temperature dependence of the recovery voltage. (a) relaxation transients parameterized by temperature (b) asymptotic recovery voltage vs. temperature.

voltage is about 600ppm, a factor of 20 times larger than in pure SiO_2. This represents an error at the 11-bit level for a converter with a 5V reference voltage. These measurements are in good agreement with the measurements of Fattaruso et al [166]. The larger recovery voltages in nitrided capacitors is not surprising based on the higher threshold voltage shifts observed in MOSFETs with nitrided oxides [Sec. 7.2.3]. The parameters required to fit the data in Fig. 7.58 are

$$\begin{aligned} D_{ot}(E) &= \begin{cases} 7.5 \times 10^{19}\text{cm}^{-3}\ \text{eV}^{-1} & V_p > 0 \\ 7.0 \times 10^{19}\text{cm}^{-3}\ \text{eV}^{-1} & V_p < 0 \end{cases} \\ \tau_{oc} &= 5.0\text{ns} \\ \tau_{oe} &= 10.0\text{ns} \\ x_o &= 0.74\mathring{A} \end{aligned}$$

The trap density for the nitrided dielectric is several hundred times greater than in the SiO_2 dielectric.

Because of the low levels of dielectric absorption measured in SiO_2 dielectrics, which border on the resolution capabilities of the test circuit, nitrided dielectrics are used as the vehicle for the remainder of the measurements in this section. The recovery voltage increases with stress time as shown in Fig. 7.59. Also shown is the theoretical result (solid line) using the same model and parameter set as in Fig. 7.58. The model also accurately predicts the observed decrease of the recovery voltage with increasing discharge time, as shown in Fig. 7.60.

The temperature dependence of the relaxation recovery waveform is shown in Fig. 7.61. The asymptotic recovery voltage exhibits a small and nearly linear increase with temperature with a coefficient of approximately $2\mu\text{V}/^\circ\text{C}$. Since the plates are degenerately doped, movement of the Fermi level with temperature is limited and the temperature dependence arises primarily from the spread of the tail of the Fermi-Dirac distribution function with increasing temperature. Thus, in an elastic tunneling model, a positive temperature coefficient can be explained by a trap distribution which increases with energy above the Fermi level. Note that mobility fluctuations do not enter the problem because there is no current flow. Using the exact model of Eqs. 5.11 and 5.16, the temperature dependence can be fit with a trap density of the same form as Eq. 7.4,

$$\begin{aligned} D_{ot}(E,x) &= \left\{7.5 \times 10^{19} + 6.0 \times 10^{20}\left[1 + e^{-(2.6-E)/0.025}\right]^{-1}\right\} \exp[-x/50.0 \times 10^{-8}] \\ \tau_{ot} &= 1.43\text{ns} \\ \tau_{ot}^{o} &= 9.1\text{ns} \\ \phi_{cb} &= 2.7\text{eV} \end{aligned} \tag{7.18}$$

In this case, since the plates are essentially metallic, it is convenient to measure the energy E downward from the oxide conduction band. This trap distribution

represents a Fermi function which increases from 7.5×10^{19} to 6.0×10^{20} cm^{-3} eV^{-1} at an energy 2.6eV below the oxide conduction band, and falls off exponentially with depth from the interface to e^{-1} of its value at the surface at a depth of $50\mathring{A}$. These theoretical results show a good fit to the data as shown in Figs. 7.61(a) and (b).

It should be noted that other mechanisms, such as lattice relaxation multiphonon emission, as discussed in Sec. 3.2.2 may also explain the positive temperature coefficient. With increasing temperature lattice vibrations increase, modulating trap energy levels so that previously inaccessible trap energies can be accessed by tunneling electrons. However, the present set of measurements is incapable of discriminating between such a mechanism and the model proposed above. Further work is required in order to determine the importance of lattice relaxation on both dielectric relaxation in capacitors as well as on threshold voltage relaxation in MOSFETs.

Chapter 8

Conclusions

8.1 Summary

This thesis has investigated the dynamics of charge trapping by tunneling to and from pre-existing oxide traps under stress conditions relevant to analog integrated circuits. In transistors this manifests itself in the form of transient threshold voltage shifts; in capacitors, as dielectric relaxation. An important difference between these stress conditions and those used in many previous studies is that (1) carriers are not injected into the oxide bands and (2) new traps are not generated during stress.

Experimentally, NMOS transistors with SiO_2 dielectrics subjected to momentary (1μs–1ms) gate-source voltage pulses of moderate amplitude (1-5V) exhibit effective threshold voltage shifts on the order of 100μV to 1mV which relax over a very long period of time (10–100ms). This problem is of concern because of its implications on the performance of MOS devices in high-accuracy analog circuit applications, for example operational amplifiers and comparators in analog-to-digital and digital-to-analog converters. The magnitude of the threshold voltage shifts in PMOS transistors are about an order of magnitude lower than in NMOS.

8.1.1 Theory

While the tunneling mechanism is certainly not new, previous experiments have overlooked the voltage and time scales of interest to analog circuits and existing tunneling theories, while assuming that electrons can be exchanged with oxide traps at different energies, have not been precise about the mechanisms by which such exchange takes place. A central theme of this thesis has been the development of a rational and consistent framework for the treatment of tunneling problems under large-signal bias conditions within which oxide band bending, temperature and trap nonuniformities in space and energy can be handled.

It was found that elastic tunneling from the conduction band alone cannot explain

the magnitude or the voltage and time dependence of the observed threshold voltage shifts. The data can be explained, however, by assuming that carriers tunnel into oxide trap energy levels which are lowered below the conduction band during stress. Various mechanisms were discussed by which such an energy loss could occur between the initial and final states in the tunneling process. Inelastic tunneling, through the assistance of phonons or other elementary excitations, is an improbable process and was neglected. However, energy may be dissipated by a thermal transition at the interface, preceeded or followed by an elastic tunneling transition (two-step process), or energy may be lost to the lattice after the electron has tunneled into an excited state of an oxide trap (lattice relaxation multiphonon emission).

A general model for the threshold voltage shift resulting from a momentary large-signal gate-source voltage pulse was derived based on these mechanisms. Time constants for each of these processes were derived using Bardeen's theory. It was found that when tunneling is the rate-limiting step, the time constant of the two-step process exhibits the expected exponential dependence on depth. Under certain conditions however, e.g. for deep traps or at low temperatures, the thermal process may become rate-limiting and the time constant becomes thermally-activated. In general the time constant lies between these two extremes. The time constant for the tunneling/lattice relaxation process is also thermally activated and depends exponentially on depth. Because of this similarity, it is difficult to distinguish between these two processes based on voltage and time measurements alone. In strong inversion, however, the two-step time constant is temperature-independent while the lattice relaxation time constant remains activated due to the barrier height ΔE_B.

Correlated mobility fluctuations, due to coulombic scattering of inversion layer charge carriers from charged oxide traps, can also contribute to an *equivalent gate-source voltage shift*. While this effect was not explicitly included in the simulations, it can be incorporated in a straightforward way. In a model in which the scattering rate is independent of trap depth and varies inversely with the total inversion charge, $S(x) = S_o/Q_N$, the trapped charge is calculated in the usual way, but is transformed into an *effective threshold voltage shift* using the substitution $t_{ox} \longrightarrow (1+\mu_N)t_{ox}$. For $x \ll t_{ox}$ the effective threshold voltage shift, which includes the effects of mobility fluctuations, is just the usual threshold voltage shift multiplied by a temperature-dependent scale factor $(1 + \mu_N S_o)$. This changes the amplitude, but not the time constant or the general form, of the solutions as a function of voltage and time.

Due to band bending in the oxide, the tunneling transitions during and after stress may occur via any combination of the aforementioned processes. In the two-step model, the three dominant combinations for an NMOS transistor were labeled *cb/cb*, *it/cb* and *it/it* and their respective dependences on stress time, relaxation time and voltage were derived. A simulation program was implemented which computes the contributions of each of these components to the total threshold voltage shift for arbitrary bias conditions and trap distributions. The generalization to include the

tunneling/lattice relaxation process was also described.

8.1.2 Experiment

Various mechanisms were suggested which could contribute to the observed threshold voltage shifts including mobile ions, dipole polarizations, and hot carriers. In order to isolate the tunneling mechanism, a test circuit was designed and implemented which enabled all of these contributions to be eliminated from the response. In addition, the following experimental observations enabled us to positively exclude thermal effects as a source of error in our measurements:

1. Shifts in drain current are observed even under conditions of constant power dissipation and in fact, even when the power dissipation *decreases* in the DUT during stress.

2. Measured gate voltage shifts are relatively insensitive to drain voltage and hence changes in power dissipation during stress.

3. Threshold voltage shifts are observed even at stress times which are much shorter ($\sim 10\mu s$) than thermal time constants.

Thus, we are able to conclude that the threshold voltage shifts measured by the test circuit are due to the tunnel exchange of inversion layer charge carriers with near-interface oxide traps. This conclusion is also supported by the high degree of correlation found between threshold voltage relaxation and 1/f noise in MOSFETs.

8.1.3 Discussion

Experimental results were presented and compared to simulations using the tunneling model. The model was found to provide a good fit over time, voltage and temperature.

NMOS

In NMOS transistors with conventional SiO_2 dielectrics, the threshold shift due to a positive (inverting) gate voltage pulse exhibits the following characteristics:

1. The threshold voltage shift increases approximately linearly with voltage.

2. The relaxation of the threshold voltage shift is logarithmic with time for $t_r \ll t_s$ but shows increasing deviations at longer times.

3. The dependence of the initial threshold voltage shift on stress time is approximately log(time) but exhibits varying amounts of a $\log^2$(time) component as well.

4. The threshold voltage shift decreases with increasing temperature.

5. Increasing the pre-stress inversion level reduces the magnitude of the threshold voltage shift.

6. The threshold voltage shift is approximately independent of drain voltage.

7. The threshold voltage shift, at constant drain current, is independent of back-gate bias.

These results can be explained by a large density of oxide traps concentrated at energies at and above the silicon conduction band. During stress, these traps are lowered below the electron quasi-Fermi level by an amount which, in strong inversion, depends linearly on the applied voltage. The deviations from strict log(time) dependence, which are not explained by earlier theories, follow naturally from the model presented here. The $\log^2$ dependence on stress time arises from the it/cb (or lr/cb) component which may be enhanced in devices with large interface state densities. The present theory predicts a logarithmic decay only at short times, approaching a $1/t$ asymptote in the long-tail regime.

The threshold voltage shifts are driven by the change, during stress, in trap energy levels with respect to the electron quasi-Fermi level at the interface. With increasing initial level of inversion, the Fermi level starts out higher in the bandgap and hence cannot change as much during stress before becoming pinned at the conduction band. This suppresses the it/it (lr/lr) component and decreases the threshold voltage shift. The positive threshold voltage shifts are nearly independent of substrate bias because the constant drain current condition keeps the position of the Fermi level constant with respect to the conduction band. The drain voltage independence is expected in the tunneling model provided that contributions due to hot carriers or thermal effects do not interfere, as guaranteed by the design of the measurement circuit and linear region operation.

The temperature dependence in strong inversion is due to three effects:

1. The movement of the Fermi "probe" Δf to lower energies with increasing temperature, coupled with a trap distribution which increases with energy above the conduction band.

2. The $T^{-3/2}$ decrease in mobility with increasing temperature, which decreases the importance of correlated mobility fluctuations relative to the threshold voltage shift. This assumes acceptor-like traps in the upper half of the bandgap and above the conduction band, which become charged during positive stress, thereby increasing coulombic scattering.

3. The activated $[\exp(\Delta E_B/kT)]$ term in the lattice relaxation process.

The first two effects cause a decrease in the equivalent gate-source voltage shift, as observed. Only factors (1) and (3) were included in the current version of the simulation program. The third effect causes an increase in threshold voltage with temperature, and required $\Delta E_B \approx 0$ in order to fit the data. It cannot be concluded from this, however, that lattice relaxation is unimportant because mobility fluctuations were not modeled, and it is possible that the positive temperature coefficient of lattice relaxation could be cancelled by the negative coefficient for correlated mobility fluctuations. Further work in this area is needed.

NMOS transistors subjected to negative gate voltage pulses show a similiar dependence on stress and relaxation time, but the voltage dependence is very different:

1. For small negative voltages there is very little threshold voltage shift, followed by an abrupt step which occurs around the point of accumulation. This step is followed by a plateau, which saturates as a function of gate voltage.

2. The gate voltage at which the step occurs shifts along the voltage axis by an amount proportional to the back gate bias.

3. The amplitude of the negative shift increases with increasing initial level of inversion.

There are two reasons for the negative threshold voltage shifts. First, since the device starts out in inversion with $V_G - V_{FB} > 0$, some of the traps above the conduction band are initially below the Fermi level and are therefore filled. During negative stress, these traps rise above the Fermi level and emit electrons, producing a negative threshold voltage shift.

In addition to this, however, is the experimental observation that the response depends on the point of accumulation and therefore on the presence of holes. We therefore conclude that the negative response is due, at least in part, to oxide traps with energies near midgap which discharge during negative stress either (1) by tunneling to an interface state followed by thermal excitation to the silicon bands (two-step) or (2) by thermal excitation of occupied oxide traps to the silicon conduction or valence band energy where elastic tunneling can occur (lattice relaxation multiphonon emission). In either case, the time constants for traps near midgap are thermally-activated and can become large when the surface carrier concentration is low. Thus, no detrapping occurs for small negative biases which drive the Fermi level near midgap. When the pulse amplitude is sufficient to accumulate the surface, oxide trap time constants are reduced and emission can occur. The abruptness of the plateau in the response is simply a consequence of the gate voltage–surface potential relationship.

These explanations are consistent with observation (2) since increasing the substrate bias shifts the voltage at which accumulation occurs, and with (3) since the higher the initial position of the Fermi level, the more traps which are initially filled

and can be emptied during a negative pulse when level sweeps down through the bandgap.

The saturation of the response observed for accumulating pulses is believed to be due to the large local electric fields created at the interface between captured majority carriers in the insulator and minority carriers in the inversion layer. This field can be sufficient for Fowler-Nordheim tunneling of the majority carriers out of oxide traps to the substrate. This detrapping occurs too rapidly to be observed by our measurements.

For negative gate voltage stress, the temperature dependence arises primarily from two sources:

1. Correlated mobility fluctuations.
   If traps in the lower half of the bandgap are donor-like, they become charged during negative stress. This increases coulombic scattering in the channel, which contributes to an increase in the amplitude of the gate-source voltage shift. Since the scattering term becomes less important at higher temperatures ($\sim T^{-3/2}$), the amplitude of the equivalent gate-source voltage shift decreases with increasing temperature.

2. The activated nature of the time constants for the two-step process and lattice relaxation.
   With decreasing temperature, trap time constants in the bandgap increase, bringing more traps within the time constant window observable in our measurements. This causes the magnitude of the gate-source voltage shift to increase with decreasing temperature.

PMOS

In PMOS devices, the threshold voltage shifts are about an order of magnitude smaller than in NMOS. In most of the devices tested, the response for negative (inverting) pulses was below the level of resolution of the measurement circuit ($< 10\mu$V). Most devices did exhibit a small shift in the positive direction, consistent with electron trapping by states within the bandgap when the device is driven into accumulation.

Several explanations may account for the much smaller threshold voltage shifts in PMOS devices:

1. The effective mass of a hole in the SiO_2 valence band is 10-20 times larger than that of an electron in the conduction band.

2. The potential barrier for holes between the SiO_2 and silicon valence bands is about 1eV higher than the barrier for electrons.

3. Many PMOS devices have buried channels, where a potential barrier physically separates inversion layer holes from the Si-SiO_2 interface.

4. PMOS devices may have lower trap densities due to processing differences.

5. The correlated mobility term in the equivalent gate-source voltage shift is larger in NMOS than in PMOS devices.

The first two items decrease the tunneling probability. However, since these quantities only enter the WKB factor as a square root, they cannot account for the complete difference. It is also unreasonable to expect large differences in the trap densities of oxides grown on different substrate types.

Recent work on $1/f$ noise supports the last explanation [118]. By Matthiessen's rule [Eq. 4.9], the larger mobility, μ_N, of electrons in NMOS devices increases the relative importance of the coulombic scattering contribution to the mobility, μ_S. Thus, correlated mobility fluctuations can dominate number fluctuations in NMOS transistors ($\mu_N S_o \gg 1$), while they are entirely negligible in PMOS ($\mu_N S_o \ll 1$). A mobility term an order of magnitude larger than the threshold voltage term could explain the much greater shift in NMOS. Further quantification of the relative magnitudes of threshold voltage (number) and mobility fluctuations are required before this theory can be verified in a quantitative way.

Nitrided Oxides

Conventional SiO_2 dielectrics were compared with nitrided and reoxidized nitrided oxides. NOX NMOS devices exhibit the same general time and voltage dependence as SiO_2 but at levels which are about an order of magnitude larger. Threshold voltage shifts in ROXNOX were comparable to, but slightly less, than in NOX. These results are consistent with 1/f noise measurements and with the higher trap densities measured in nitrided oxides and their slight reduction by reoxidation. PMOS devices in these processes, in contrast to SiO_2 exhibit threshold voltage shifts which are comparable to NMOS. This is not consistent with 1/f noise measurements [118] and warrants further study. However, since 1/f noise probes only trap energies in the immediate vicinity of the minority carrier quasi-Fermi level, while threshold voltage measurements probe a much broader range, there is no reason to expect perfect correlation. The time dependence of the PMOS relaxations are the same as in NMOS but the voltage dependence is of the reverse polarity, increasing monotonically with negative gate voltage pulses and saturating in the positive direction.

Dielectric Relaxation

In the case of MOSFETs, we began with the experimental observation of a new phenomenon, then developed theory in order to explain it. In the capacitor study, this procedure was reversed. Based on the tunneling theory, we derived the functional form of the recovery voltage that should develop across a capacitor when it is subjected to voltage stress for a period of time, momentarily discharged and then open-circuited.

Measurements were then taken in order to verify the theory. The tunneling model was shown to provide a good fit to the data as a function of stress voltage, stress time, relaxation time and discharge time. The voltage dependence was found to be linear while the time dependence obeys a $\log^3(\text{time})$ law. Capacitors with nitrided dielectrics were shown to have higher levels of recovery voltage than those with pure SiO_2, consistent with higher trap densities as expected from the transitor study. The trap densities extracted on the basis of dielectric relaxation measurements were of the same general form as those extracted from threshold voltage measurements. The magnitudes were somewhat higher, which can be attributed to the fact that the capacitor dielectrics were deposited rather than thermally grown.

It can be concluded that interfacial tunneling provides a plausible explanation for dielectric relaxation in integrated circuit capacitors on the time and voltage scales of interest for analog circuits.

Circuit Implications

It was shown that threshold voltage shifts due to accumulating pulses (negative in NMOS and positive in PMOS) could be greatly reduced through the application of a backgate bias. Shifts for inverting pulses can also be suppressed, to a lesser extent, by strong inversion operation. It was also shown that PMOS transistors with SiO_2 dielectrics exhibit very low threshold voltage shifts, and the shifts which are observed are primarily in the positive direction. These considerations enable threshold voltage hysteresis effects to be greatly reduced in circuits. By operating a PMOS transistor with a backgate bias higher than the largest expected gate voltage pulse, the threshold relaxation effect can be virtually eliminated.

A simple circuit model was presented which captures the general features of the threshold voltage relaxation for the purpose of circuit simulations. This model can be used in SPICE to predict the effect of threshold voltage hysteresis effects on circuit performance.

Errors due to threshold voltage relaxation were estimated to have an important effect on the performance of analog-to-digital converters at accuracies above the 12-bit level.

8.2 Suggestions for Future Work

The focus of this thesis has been on physically explaining and quantitatively modeling relaxation effects that occur in MOS transistors and capacitors under large-signal transient stress conditions in circuits. The results of this work have broader applications and suggest additional experiments and theory that would enhance our understanding of relaxation effects and of tunnel exchange with oxide traps in general. We have already pointed out applications to the interpretation of DLTS and

charge pumping measurements as well as to threshold voltage instabilities in II-VI and III-V MISFETs. Some other potential areas for future research are suggested below.

Circuit-Level Modeling

The important implications of trapping–induced relaxation effects on the performance of high–accuracy circuits have been emphasized throughout this thesis. We have presented a detailed device–level model as a means to verify our proposed physical mechanism and the model has adequately served this purpose. However there are other areas where the results of this research are applicable but a different type of model implementation is required. One example of this is in circuit-level simulation, another is in the measurement of trap densities as described below.

A simplified version of the model is desirable for the simulation of trapping–induced threshold voltage shifts during the integrated circuit design process. In a circuit simulation environment, efficiency rather than rigorous physical correctness, is the primary concern. Progress has been made in the form of the SPICE subcircuit model described in this thesis. However, more efficient models are required which can be embedded into the simulator's model routines rather than in the form of a subcircuit. The development of such a model should be greatly facilitated by the improvements in our physical understanding provided by this thesis.

Theory

The work presented in this thesis also suggests new areas for theoretical research. This work has presented a theoretical framework which is complete, in the sense that it contains all of the essential elements necessary to describe the observed relaxation phenomena. Nevertheless, certain aspects of the model have not been fully quantified; in particular, the relative importance of (1) lattice relaxation versus the two-step process, and (2) mobility versus threshold voltage fluctuations.

We have shown that both the lattice relaxation and two-step models can explain the voltage and time dependence of the experiments. Their primary difference is in the strong inversion temperature dependence. However, in order to accurately model the temperature dependence, the second issue must be addressed, since the correlated mobility fluctuations depend on temperature. Further work is required to determine the relative importance of the effect of a trapped charge carrier in producing a threshold voltage shift and in increasing coulombic scattering in the channel. While the answers to these questions will not change any of the general results presented in this thesis, they will influence the interpretation of simulation results, e.g. the trap densities extracted from relaxation measurements.

Temperature dependence is the key to understanding these issues. Dielectric relaxation measurements on capacitors present a promising direction for future research

because there is no current flow and the mobility fluctuation term drops out. By comparing threshold voltage relaxation in transistors and dielectric relaxation in capacitors with identically grown dielectrics, the importance of the mobility fluctuation term may be determined.

The resolution of these issues may well lie in alternative measurement techniques. Large-signal relaxation measurements are inherently difficult to interpret because they involve large excursions of the Fermi level across the bandgap, bringing into play a broad range of traps and a large number of mechanisms. A technique which isolates one feature at a time is better suited to resolve detailed questions of mechanism. One promising technique is 1/f noise, particularly on small-geometry MOSFETs where trapping by individual defects can be observed [116, 118].

As in the case of 1/f noise, the very long relaxation times observed in this thesis depend on the existence of a wide distribution of time constants spanning many decades. The present approach, following in the tradition of McWhorter [2], has obtained this distribution of time constants from the exponential dependence of the tunneling time constant on spatial depth. The lattice relaxation mechanism suggests an entirely different approach, where the time constant dispersion is obtained from a broad distribution of activation energies ΔE_B for thermal capture. This approach is worthy of further investigation.

Finally, further research is required to model the very fast emission that occurs under gate voltage pulses that drive the surface into accumulation and then return it to inversion. We have postulated that this effect is due to field emission under the high electric fields created by trapped majority carriers and inversion layer minority carriers separated by a thin insulating layer. These high field effects were not quantitatively modeled in this thesis, and further research is needed in order to confirm this theory.

Measurement of Oxide Trap Densities

The opportunities for using dielectric relaxation and transient threshold voltage measurements as a technique for probing oxide trap densities have only begun to be explored in this thesis. Since the time constant of an oxide trap depends exponentially on the depth of a trap, measurements of threshold voltage shift or recovery voltage with stress time contain information on the spatial distribution of traps. Similarly, since the range of trap energies which contribute to the voltage response depends on the magnitude of the applied gate voltage pulse, measurements of the response as a function of stress voltage contain information about the energy distribution of the traps.

We have extracted rough estimates of the trap energy distributions in this thesis but a more detailed determination will require a more efficient simulator. Such a simulator must be physically correct but fast enough so that optimized fits can be produced for a large quantity of data. One promising approach is to replace

the numerical integrations with a finite element method in the spirit of Sec. 3.5.4, discretizing the equations in the three dimensions of time, space and energy.

Accurate trap density measurements will also require precise knowledge of the relative magnitudes of the mobility and threshold voltage terms in the equivalent gate-source voltage shift, since this affects the scale factor that multiplies the trap density.

Processing Solutions

It has been shown that bias conditions can be used to suppress trapping-induced threshold voltage shifts under certain conditions, but the key to improving the large-signal transient performance of MOS devices in analog circuits lies in improving processing techniques to produce cleaner interfaces. We have shown that relaxation effects are directly related to trap densities and have demonstrated large variations between processes. Further research is required in order to correlate processing variations, e.g. oxidation conditions, with measured threshold voltage shifts and recovery voltages. The production of higher accuracy MOS analog circuits, both from the point of view of large-signal relaxation effects as well as 1/f noise, depends on the elimination of electron and hole traps at the interface and in the dielectric.

Appendix A

Tunneling Transition Probability

A.1 Introduction

In this appendix we seek the answer to the following fundamental question:

> Given an electron initially localized on one side of a potential barrier (in a conduction band, valence band or interface state), what is the probability, per unit time, of its making a transition to a state localized on the other side of the barrier (an oxide trap)?

The problem in which we are interested involves the tunneling of an electron between a semiconductor and an oxide trap, which always occurs through a trapezoidal barrier and across a region of discontinuous band structure. Surprisingly, this problem has been addressed by relatively few researchers. In contrast, the problem of field ionization of electrons trapped in localized potential wells in insulators and semiconductors has been treated extensively [217, 218, 219, 220, 101, 221, 222, 223, 224, 225]. Field ionization of traps can occur under high fields when the insulator bands bend sufficiently for an electron in a bound state to tunnel through the triangular barrier in which it is localized into the oxide conduction band. The high fields required for this condition will generally not occur during our measurements or in analog circuits under normal operating conditions. A possible exception to this statement may occur when the surface of a MOSFET is pulsed from accumulation into strong inversion, as discussed in Sec. 7.2.2.

The tunneling problem can be rigorously formulated as a time dependent initial value problem, in which the electron is localized on one side of the barrier, by constructing a wavepacket from eigenstates of the exact one-electron Hamiltonian. The integrated probability density will then decrease with time on one side of the barrier, and increase on the other, corresponding to the electron tunneling through the forbidden region. This approach is demonstrated in the case of the two-step process in Appendix B.2.

In the case of tunneling to oxide traps from band states, this method is mathematically complex and the wavepacket can never be completely localized on one side of the barrier if constructed from exact eigenstates of the Hamiltonian. Oppenheimer [226] in his analysis of the field ionization of atomic hydrogen, overcame these problems by using an approximate method in which the initial and final states were taken to be stationary states of different Hamiltonians. In the hydrogen problem, the initial state was taken to be a bound state of an electron in a Coulomb potential and the final state an electron in an electric field.

Bardeen [227] later extended this method to MIM tunneling junctions. Since the tunneling probability is very small, the barrier separates the system into two nearly independent portions. The initial state is taken to be an electron incident on an insulator extending infinitely to the right, while the final state consists of an insulator extending infinitely to the left. The solution to the exact Hamiltonian is written as a linear combination of these basis states and, when substituted into the time-dependent Schrödinger equation, yields an expression for the transition probablity resembling Fermi's Golden Rule of time dependent perturbation theory. Instead of a time dependent perturbation in the matrix element, one obtains a *transfer Hamiltonian* which weakly couples the two sides of the junction and is interpreted as transfering electrons from one side to the other. This method, which has come to be known as the *transfer Hamiltonian method*, has been used successfully to describe gap structures in the current-voltage characteristics of tunnel junctions with superconducting electrodes, the Josephson effects and phonon structure in MIM junctions [228]. The transfer Hamiltonian method, in the form described by Bardeen, is reviewed below and applied to the calculation of time constants for tunneling to and from oxide traps in Appendix B.

A.2 The Transfer Hamiltonian Formalism

In time dependent perturbation theory, the exact Hamiltonian is written in the form

$$H = H_o + H_1(t) \tag{A.1}$$

where the solutions to the time-independent part H_o are known and $H_1(t)$ is a small time-dependent perturbation. The transition rate from an initial state ψ_i to a continuum of final states ψ_f, where ψ_i and ψ_f are eigenfunctions of H_o, is given by Fermi's golden rule [229],

$$W_{if} = \frac{2\pi}{\hbar}|\langle\psi_f|H_1|\psi_i\rangle|^2\rho(E_f)\delta(E_i - E_f) \tag{A.2}$$

where $\rho(E_f)$ is the density of final states. Since we are interested in calculating the transitition rate from bound or band states in the semiconductor to oxide traps, it is natural to turn to time dependent perturbation theory. The question therefore arises

of how separate out of the total Hamiltonian a small perturbation term, which can be used in Eq. A.2.

Bardeen's approach [227, 131, 230], based on the original formulation of Oppenheimer [226, 231], consists of writing the exact Hamiltonian in terms of two partial Hamiltonians, each of which is exact in overlapping regions of configuration space. In the case of a planar geometry, such as an MIS structure, with x the direction perpendicular to the plane, the model Hamiltonian is

$$H = \begin{cases} H_l & x < x_o \\ H_r & x > 0 \end{cases}$$

It is assumed that the solutions to these two partial Hamiltonians are known,

$$\begin{aligned} H_l\psi_l &= E_l\psi_l \\ H_r\psi_n &= E_n\psi_n \qquad n \neq l \end{aligned}$$

The system is assumed to be initially in the state ψ_l at $t = 0$ and we seek the transition rate into a a set of final states ψ_n. For example, in the problem of tunneling to an oxide trap, ψ_l is either a Bloch state in the conduction band or a bound state of an interface potential well . The final state ψ_n is a bound state of an oxide trap, as shown in Figs. B.2 and B.7.

Since H_l and H_r are different Hamiltonians, the direct product of their eigenfunctions do not necessarily form a complete basis set for the eigenfunctions of H. Nevertheless, it is assumed that the eigenfunctions of H_l and H_r are nearly orthogonal and that their direct product is a good approximation to a complete set of states. This assumption is justified on the basis that tunneling only weakly couples the two sides of the barrier so that they can be regarded as nearly independent systems. The trial solutions of H are then expanded as

$$\psi(r,t) = \psi_l e^{-iE_lt/\hbar} + \sum_{n\neq l} a_n(t)\psi_n e^{-iE_nt/\hbar} \tag{A.3}$$

where $a_n(t = 0) = 0$ so that $\psi(r, 0) = \psi_l$.

It should be noted that the major difference between time-dependent perturbation theory in its usual form, and in the form proposed by Oppenheimer, is that in the former, the basis states used in the expansion of the exact solution, Eq. A.3, are exact eigenstates of an approximate Hamiltonian (H_o), whereas in the latter, the basis consists of approximate eigenstates (ψ_l, ψ_n) of the exact Hamiltonian (H).

The remainder of the analysis follows the usual derivation of time-dependent perturbation theory. The time dependent coefficients $a_n(t)$ are found by substituting the assumed form of the solution Eq. A.3 into the exact Schrödinger equation

$$H\psi = i\hbar\frac{\partial\psi}{\partial t} \tag{A.4}$$

to give

$$E_l\psi_l e^{-iE_lt/\hbar} + \sum_{n\neq l}\left(a_nE_n\psi_n e^{-iE_nt/\hbar} + i\hbar\dot{a_n}\psi_n e^{-iE_nt/\hbar}\right) = H\psi_l e^{-iE_lt/\hbar} + \sum_{n\neq l} a_nE_nH\psi_n e^{-iE_nt/\hbar} \quad (A.5)$$

For small t, $a_n(t) \approx 0$ for $n \neq l$ since the electron is initially in state ψ_l. Multiplying both sides by $\psi_r^* e^{iE_rt/\hbar}$ and integrating over all space,

$$\begin{aligned}\int_{-\infty}^{+\infty}\psi_r{}^*(H-E_l)\psi_l e^{-i(E_l-E_r)t/\hbar}d^3r &= i\hbar\sum_{n\neq l}\dot{a}_n\int_{-\infty}^{+\infty}\psi_r{}^*\psi_n e^{-i(E_n-E_r)t/\hbar}d^3r \\ &= i\hbar\dot{a}_r(t) \end{aligned} \quad (A.6)$$

where the wavefunctions are assumed orthonormal. Thus,

$$\begin{aligned}\dot{a}_r(t) &= -\frac{i}{\hbar}\int_{-\infty}^{+\infty}\psi_r{}^*(H-E_l)\psi_l e^{-i(E_l-E_r)t/\hbar}d^3r \\ &= -\frac{i}{\hbar}M_{lr}e^{i\omega_{lr}t} \end{aligned} \quad (A.7)$$

where $\omega_{lr} = (E_r - E_l)/\hbar$, and

$$M_{lr} = \int\psi_r{}^*(H-E_l)\psi_l d^3r = \langle\psi_r|H-H_l|\psi_l\rangle \quad (A.8)$$

is the effective matrix element for tunneling, as can be seen by integrating Eq. A.7,

$$a_r(t) = -\frac{i}{\hbar}\int_0^t M_{lr}e^{i\omega_{lr}t'}dt' = \frac{M_{lr}}{i\hbar}\frac{(e^{i\omega_{lr}t}-1)}{i\omega_{lr}} \quad (A.9)$$

The probability of finding the electron in the state r after a time t is then

$$|a_r(t)|^2 = \frac{4|M_{lr}|^2}{\hbar^2}\frac{\sin^2(\omega_{lr}t/2)}{{\omega_{lr}}^2} \quad (A.10)$$

The total transition rate, per unit time, out of the state ψ_l to a group of final states of density $\rho_r(E_r)$ is

$$W_{lr} = \frac{4|M_{lr}|^2}{\hbar^2 t}\int_{-\infty}^{\infty}\frac{\sin^2(\omega_{lr}t/2)}{\omega_{lr}^2}\rho_r(E_r)dE_r \quad (A.11)$$

Since the integrand is sharply peaked at $\omega_{lr} = 0$, the density of states can be evaluated at $E_l = E_r$ and removed from the integral, which is easily evaluated [229] to give, for the transition rate per unit time out of the state ψ_l,

$$W_{lr} = \frac{2\pi}{\hbar}|M_{lr}|^2\rho_r(E_r)\delta(E_r - E_l) \quad (A.12)$$

which is just Fermi's golden rule. Bardeen's method provides an explicit formula (Eq. A.8) for the evaluation of the matrix element in Eq. A.12 in the case of tunneling through an insulating barrier.

Note that, if the Hamiltonian is written in the form of Eq. A.1, with ψ_l an eigenstate of H_o and H_1 a perturbation, then,

$$M_{lr} = \langle \psi_r | H - E_l | \psi_l \rangle = \langle \psi_r | H - H_o | \psi_l \rangle = \langle \psi_r | H_1 | \psi_l \rangle \tag{A.13}$$

which resembles the transition matrix element in ordinary time dependent perturbation theory. There is, however, an important difference between the matrix elements in Eqs. A.2 and A.12. In Eq. A.2 the matrix element is calculated between eigenstates of the unperturbed Hamiltonian H_o whereas in Bardeen's approximation, Eq. A.12, the matrix element is taken between eigenstates of different Hamiltonians.

The tunneling matrix element can be rewritten in a more elegant form as shown by Bardeen, by rewriting Eq. A.8 as,

$$\begin{aligned} M_{lr} &= \int_{-\infty}^{+\infty} \psi_r^*(H - E_l)\psi_l d^3r \\ &= \int_{-\infty}^{+\infty} dz \int_{-\infty}^{+\infty} dy \int_{x_B}^{\infty} [\psi_r^*(H - E_l)\psi_l - \psi_l(H - E_r)\psi_r^*]\, dx \end{aligned} \tag{A.14}$$

where $0 < x_B < x_o$, since $(H - E_l)\psi_l = 0$ for $x \leq x_o$ and $(H - E_r)\psi_r = 0$ for $x \geq 0$. In the one-dimensional potentials of interest, the exact Hamiltonian can be written as the sum of kinetic and potential energy,

$$H = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V(x) \tag{A.15}$$

Thus,

$$\begin{aligned} M_{lr} = &-\frac{\hbar^2}{2m}\int_{-\infty}^{+\infty} dz \int_{-\infty}^{+\infty} dy \int_{x_B}^{\infty} \left[\psi_r^* \frac{d^2\psi_l}{dx^2} - \psi_l \frac{d^2\psi_r^*}{dx^2}\right] dx \\ &+ \int_{-\infty}^{+\infty} dz \int_{-\infty}^{+\infty} dy \int_{x_B}^{\infty} [\psi_r^* V(x)\psi_l - \psi_l V(x)\psi_r^*]\, dx \\ &- \int_{-\infty}^{+\infty} dz \int_{-\infty}^{+\infty} dy \int_{x_B}^{\infty} [E_l\psi_r^*\psi_l + E_r\psi_l\psi_r^*]\, dx \end{aligned}$$

The second term vanishes because the potential $V(x)$ is Hermitian and the third term vanishes because the tunneling is elastic, for times long enough not to violate Heisenberg's uncertainty principle, so that the delta function in Eq. A.12 is valid. Integrating the first term by parts, with the boundary condition that both wavefunctions vanish as x approaches infinity,

$$M_{lr} = -\frac{\hbar^2}{2m}\int_{-\infty}^{+\infty} dz \int_{-\infty}^{+\infty} dy \left[\psi_r^* \frac{d\psi_l}{dx} - \psi_l \frac{d\psi_r^*}{dx}\right]_{x=x_B} \tag{A.16}$$

Thus, the tunneling matrix element is just the matrix element of the quantum mechanical current density operator, evaluated at any point $0 < x_B < x_o$ in the barrier where both partial Hamiltonians are exact.

Bardeen's method answers the question posed at the beginning of this section. The "perturbation" which enters Fermi's golden rule and that drives the coherent, elastic contribution to the tunneling current through the average potential barrier is given by $H_1 = H - E_l$. This is frequently referred to as a "transfer Hamiltonian", an appellation which becomes more meaningful in the many-body theory of tunneling, since it transfers electrons from one side of the barrier to the other. Although $H - E_l$ is time independent, it can be thought of as a time-dependent, constant perturbation which is turned on at $t = 0$. Any explicit time dependence in the barrier Hamiltonian, e.g. potential fluctuations due to phonons or molecular vibrations, are treated as perturbations in the usual time-dependent perturbation theory and lead to inelastic contributions to the tunneling current [231].

Appendix B

Tunneling Time Constants

This appendix contains the derivation of the time constants for tunneling to oxide traps from conduction band states, in Sec. B.1 and from interface states in Sec. B.2. An exact solution, including the dependence of the time constant on electric field, energy and space, is mathematically cumbersome and lends little insight. The analyses in Sections B.1 and B.2 are therefore each broken up into two steps. First, a simple one-dimensional approximation for the trap potential is made and the WKB approximation is used for the wavefunctions. We assume a trapezoidal barrier and image force barrier lowering is intentionly not included in the analysis, since it has been shown that the classical image force barrier lowering does not contribute to tunneling. Weinberg et al [232] have found in Fowler-Nordheim tunneling experiments that their data can be explained using a model that does not include the image force but cannot be fit to a model in which the barrier lowering is included. The reasons for this are still poorly understood, but indicate that the classical image force is not applicable to tunneling and that a rigorous quantum mechanical formulation is still needed.

The main purpose of this analysis is to determine the dependence of the time constant on the electric field in the oxide. This effect would be important in creating hysteresis effects if, for example, the electric field caused the time constant during stress to be less than during emission. Lakhardi et al [73] have suggested such a field-assisted tunneling effect in order to explain an asymmetry between the time constants extracted for capture and emission in DLTS measurements. The electric field is shown to affect primarily the exponential part while the prefactor has only a very weak field dependence.

While the WKB approximation allows the effect of the electric field to be incorporated in the analysis, its accuracy near the interface is questionable. Recall that the validity of the WKB approximation requires [233],

$$\left|\frac{d\lambda}{\lambda}\right| = \left|\frac{mh}{p^3}\frac{dV}{dx}\right| \ll 1 \tag{B.1}$$

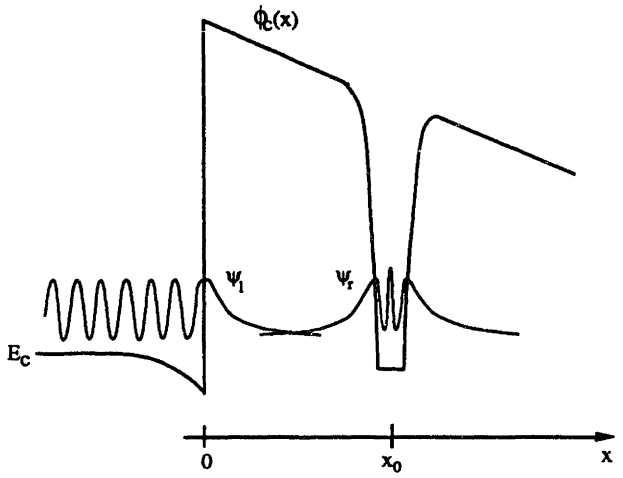


Figure B.1: Energy band diagram showing an electron tunneling from a conduction band state, on the left side of the insulator potential barrier, into an oxide trap on the right.

That is, the potential energy must change slowly enough that the momentum of the particle can be regarded as constant over many wavelengths. This condition is violated near the interface, which is the region in which we are most interested, because the potential changes abruptly and dV/dx is large. Furthermore, most of the tunneling occurs near the band edges where the electron momentum $p = \hbar k$ is near zero and the wavelength is very large. The WKB approximation is therefore abandoned in the second part of the analysis in favor of a simple rectangular barrier with matching of the wavefunctions and their derivatives across the interface. A more realistic, three-dimensional, model for the oxide trap potential is also employed. This results in a more accurate expression for the prefactor as a function of trap energy and depth. We then combine the WKB factor from the first part of the analysis with the prefactor from the second and take this as the tunneling time constant in the model.

B.1 Band-to-Trap Tunneling

The band diagram for an electron tunneling from a conduction band state on the left side of an insulating barrier to an oxide trap on the right side is shown in Fig B.1. The exact Schrödinger equation is,

$$\left[-\frac{\hbar^2\nabla^2}{2m^*} + U_c(x)\theta(x) + V_t(r) - E\right]\psi = 0$$

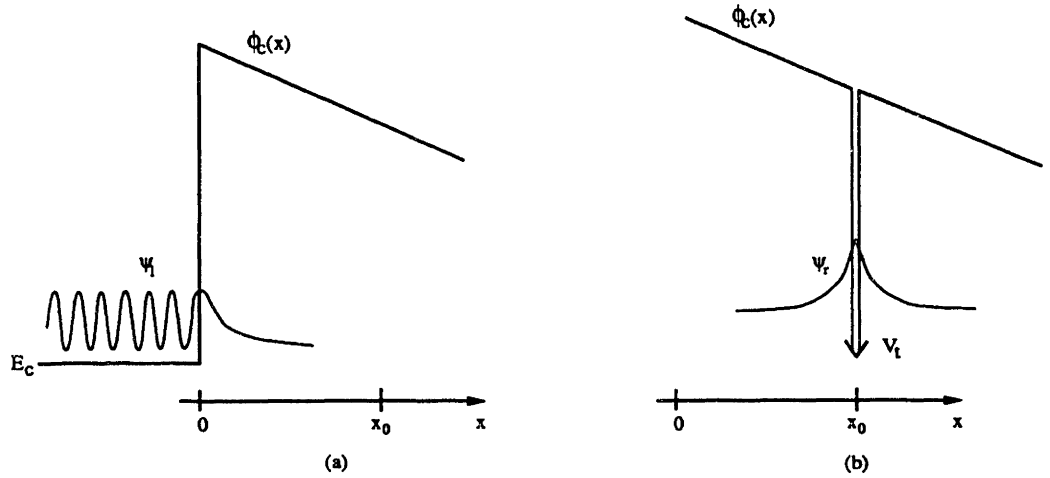


Figure B.2: Left and right side systems for use in Bardeen's method

where the barrier potential is

$$U_c(x) = \phi_c + q\mathcal{E}x \tag{B.2}$$

ϕ_c is the height of the barrier between the silicon and oxide conduction bands (approximately 3.1 eV), E is the electric field in the insulator and q is the magnitude of the electronic charge. The theta function is defined as,

$$\theta(x) = \begin{cases} 0 & x < 0 \\ 1 & x > 0 \end{cases}$$

In the effective mass approximation,

$$m^* = \begin{cases} m_l & x < 0 \\ m_r & x > 0 \end{cases}$$

In Bardeen's method, the exact Hamiltonian is separated into two partial Hamiltonians, the eigenfunctions of which represent the initial and final states of the system. Thus,

$$H_l\psi_l = E_l\psi_l \tag{B.3}$$
$$H_r\psi_r = E_r\psi_r \tag{B.4}$$

where $E_l = E_r = E$ since the tunneling is elastic. The left side wavefunction, ψ_l, is taken to be a free particle of effective mass m_l incident on a barrier extending infinitely to the right as shown in Fig. B.2(a). The Hamiltonian for this system is,

$$H_l = -\frac{\hbar^2\nabla^2}{2m^*} + U_c(x)\theta(x) \tag{B.5}$$

The right side wavefunction, ψ_r, is a bound state of an oxide trap potential well in an insulator of infinite thickness, described by the Hamiltonian,

$$H_r = -\frac{\hbar^2\nabla^2}{2m^*} + U_c(x) + V_t(r) \tag{B.6}$$

B.1.1 One-dimensional Oxide Trap Potential

We begin with a trapezoidal potential barrier and a simple one-dimensional delta function model for the oxide trap,

$$U_c(x) = \phi_c + q\mathcal{E}x \tag{B.7}$$
$$V_t(r) = V_\delta\delta(x - x_o) \tag{B.8}$$

The trap is assumed to cover the entire plane at $x = x_o$, an assumption that will be modified later. This model is obviously very crude, but the time constant will be seen to be dominated by the exponential spatial dependence arising from the overlap of the wavefunctions in Eq. A.8, which is not very sensitive to the details of the model.

We assume the specular boundary condition, in which the transverse component of the electrons wavevector, $\vec{k}_{||}$ is conserved across the interface during tunneling. The following analysis parallels that of Freeman [146], with minor modifications. This will not always be the case; for example $\vec{k}_{||}$ is not conserved in phonon-assisted inelastic tunneling, or even in elastic processes if there are fluctuations in the average barrier potential due to impurities or variations in thickness [231]. In elastic tunneling processes however, the specular boundary condition is normally assumed.

Since the potential depends only on x, the solutions for both ψ_l and ψ_r are of the form

$$\psi_{l,r}(\vec{r}) = e^{i(k_y y + k_z z)}\chi_{l,r}(x) = e^{\vec{k}_{||}\cdot\vec{r}_{||}}\chi_{l,r}(x) \tag{B.9}$$

where $\vec{r}_{||} = (0, y, z)$ and $\vec{k}_{||} = (0, k_y, k_z)$ are the components of the position vector and wavevector in the plane of the junction. Substituting Eqs. B.5 and B.9 into Eq. B.3 yields an equation for $\chi_l(x)$,

$$\left[-\frac{\hbar^2}{2m^*}\frac{d^2}{dx^2} + U_c(x)\theta(x) - E_x\right]\chi_l(x) = 0 \tag{B.10}$$

where

$$E_x = E - \frac{\hbar^2\vec{k}_{||}^{\,2}}{2m_l} = E - E_{||} \tag{B.11}$$

and $E_l = E$ is the energy of the incident electron. All energies are referenced to the bottom of the silicon conduction band at the interface.

In the WKB approximation, the solutions to Eq. B.10 are [233]

$$\chi_l(x) = \begin{cases} \frac{2C_l}{\sqrt{k_x}} \cos\left(\int_x^0 k_x(x')dx' - \frac{\pi}{4}\right) & x < 0 \\ \frac{C_l}{\sqrt{K_x}} \exp\left(-\int_0^x K_x(x')dx'\right) & x > 0 \end{cases} \tag{B.12}$$

with

$$K_x^2 = \frac{2m_r}{\hbar^2}[U_c(x) - E_x] \tag{B.13}$$

$$k_x^2 = \frac{2m_l}{\hbar^2}E_x \tag{B.14}$$

The energy of the conduction band, and hence E_x, vary with position due to band bending in the semiconductor. For simplicity, we neglect this effect as it affects only the normalization of ψ_l and would greatly complicate the analysis. Thus, $k_x(x) = k_x(0) = k_x$ and $E_x(x) = E_x(0) = E_x$.

Since the probability of finding the electron in the barrier is small, the wavefunction is normalized in the semiconductor region only, which is assumed to have dimensions L_x, L_y and L_z and volume $\Omega = L_xL_yL_z$. This yields

$$C_l = \left(\frac{k_x}{2\Omega}\right)^{1/2} \tag{B.15}$$

The solutions for ψ_r are also of the form Eq. B.9. Substituting the assumed form of the solution, together with Eq. B.6 into Schrödinger's equation, B.4, the WKB trap wavefunction χ_r is,

$$\chi_r(x) = \begin{cases} \frac{C_r}{\sqrt{K_x(x)}} \exp\left[-\int_x^{x_o} K_x(x')dx'\right] & x < x_o \\ \frac{C_r}{\sqrt{K_x(x)}} \exp\left[-\int_{x_o}^x K_x(x')dx'\right] & x > x_o \end{cases} \tag{B.16}$$

Normalizing the trap wavefunction in the insulator, assuming $K_x(x) \approx K_x(x_o)$ since the wavefunction decays exponentially away from $x = x_o$,

$$C_r = \frac{K_x(x_o)}{A^{1/2}} \tag{B.17}$$

with $A = L_yL_z$ the area of the interface.

Substituting the left and right side wavefunctions, Eqs. B.16 and B.12 into the equation for the the matrix element, Eq. A.16 yields,

$$M_{lr} = \frac{\hbar^2 A}{m_r} C_r^* C_l \exp\left[-\int_0^{x_o} K_x(x')dx'\right] \tag{B.18}$$

$$= \frac{\hbar^2 K_x(x_o)}{m_r}\left(\frac{k_x}{2L_x}\right)^{1/2} \exp\left[-\int_0^{x_o} K_x(x')dx'\right] \tag{B.19}$$

Recall from Eq. 3.10 that the net rate of change of the density of occupied oxide traps $n_{ot} = f_r D_{ot}$ at depth x and energy E is governed by the differential equation,

$$r = \frac{\partial n_{ot}}{\partial t} = (f_l - f_r) D_{ot}(E, x)\omega \tag{B.20}$$

where $\omega = 1/\tau$ is the tunneling rate. However, the one-dimensional model used for the trap assumes that the trap covers the entire plane at x. Thus, all of the electrons which manage to tunnel to x will surely be captured. Following Freeman [146], we assume that the traps have a capture cross section σ_n so that the actual rate at which electrons are captured is $\frac{\sigma_n}{A} r$. The cross section can be incorporated into the tunneling rate ω so that Eq. 3.17 becomes

$$\omega(E, x) = \frac{2\pi\sigma_n}{\hbar A} \sum_{\vec{k}_{||}} \rho_l(E) |M_{lr}|^2 \tag{B.21}$$

$$= \frac{L_x \sigma_n m_l}{\pi \hbar^3} \int \frac{|M_{lr}|^2}{k_x} \vec{k}_{||} d\vec{k}_{||} \tag{B.22}$$

Substituting Eq. B.19 in Eq. B.22,

$$\omega(E, x) = \frac{\hbar \sigma_n m_l}{2\pi {m_r}^2} \int_0^{\vec{k}_{||_{max}}} [K_x(x)]^2 \exp\left[-2\int_0^x K_x(x')dx'\right] \vec{k}_{||} d\vec{k}_{||} \tag{B.23}$$

where $K_x(x)$ is a function of $\vec{k}_{||}$, since, from Eq. B.13 and the conservation of $\vec{k}_{||}$ across the interface,

$$K_x(x)^2 = \frac{2m_r}{\hbar^2}[U_c(x) - E_x]$$

$$= \frac{2m_r}{\hbar^2}\left[U_c(x) - \left(E - \frac{\hbar^2 \vec{k}_{||}^{\,2}}{2m_r}\right)\right]$$

$$= K^2 + \vec{k}_{||}^{\,2} \tag{B.24}$$

where

$$K^2 = \frac{2m_r}{\hbar^2}[U_c(x) - E] \tag{B.25}$$

$$k^2 = k_x^2 + k_{||}^2 = \frac{2m_l E}{\hbar^2} \tag{B.26}$$

Since $E \ll U_c(x)$ near the interface at the fields of interest, it follows from Eqs. B.25 and B.26 that $k_{||}^2 \leq k^2 \ll (m_l/m_r)K^2$. Since $m_l \approx 0.1 m_o$ and $m_r \approx 0.4 m_o$ [234], it also follows that

$$k_{||}^2 \leq k^2 \ll K^2 \tag{B.27}$$

so that Eq. B.24 can be expanded as,

$$K_x \approx K + \frac{\vec{k}_{||}^{\,2}}{2K} \tag{B.28}$$

thus

$$\omega = \frac{\hbar\sigma_n m_l}{2\pi {m_r}^2} e^{-2\int_0^x K(x')dx'} \int_0^{\vec{k}_{||_{max}}} [K(x)^2 + \vec{k}_{||}^{\,2}] e^{-\vec{k}_{||}^{\,2}\int_0^x \frac{dx'}{K(x')}} d\vec{k}_{||} \tag{B.29}$$

Since the integrand is sharply peaked around $\vec{k}_{||} = 0$, the slowly varying part of the integrand, $(K(x)^2 + \vec{k}_{||}^{\,2})$ can be evaluated at $\vec{k}_{||} = 0$ and taken outside of the integral. With $\vec{k}_{||_{max}} = k = \sqrt{2m_l E/\hbar^2}$ the integral can be evaluated to give,

$$\omega(E,x) = \frac{\hbar\sigma_n K(x)^2 m_l}{4\pi {m_r}^2} \left[\frac{1 - \exp\left(-k^2 \int_0^x \frac{dx'}{K(x')}\right)}{\int_0^x \frac{dx'}{K(x')}}\right] \exp\left(-2\int_0^x K(x')dx'\right) \tag{B.30}$$

Substituting $U_c(x)$ from Eq. B.2 together with Eq. B.25 into Eq. B.30 and carrying out the integrations, the tunneling time constant is

$$\tau_{ot}(E,x) = \tau_o(E,x) \exp\left[-\frac{4}{3}\left(\frac{2m_r}{\hbar^2}\right)^{1/2} \frac{(\phi_c - E + q\mathcal{E}x)^{3/2} - (\phi_c - E)^{3/2}}{q\mathcal{E}}\right] \tag{B.31}$$

where

$$\begin{aligned}
\frac{1}{\tau_o(E,x)} &= \frac{q\sigma_n\mathcal{E}}{4\pi\hbar^2}\left(\frac{2{m_l}^2}{m_r}\right)^{1/2}\left[\frac{\phi_c - E + q\mathcal{E}x}{(\phi_c - E + q\mathcal{E}x)^{1/2} - (\phi_c - E)^{1/2}}\right] \qquad \text{(B.32)} \\
&\quad \times \left\{1 - \exp\left[-\frac{2E}{\hbar}\left(\frac{2{m_l}^2}{m_r}\right)^{1/2} \frac{(\phi_c - E + q\mathcal{E}x)^{1/2} - (\phi_c - E)^{1/2}}{q\mathcal{E}}\right]\right\} \\
&= \frac{q\sigma_n\mathcal{E}}{4\pi\hbar}\left(\frac{m_l}{m_r}\right)^2 \left[\frac{K^2(x)}{K(x) - K(0)}\right] \\
&\quad \times \left\{1 - \exp\left[-\frac{\hbar^2 m_l k^2 (K(x) - K(0))}{{m_r}^2 q\mathcal{E}}\right]\right\} \qquad \text{(B.33)}
\end{aligned}$$

The prefactor can be seen to depend on energy, electric field and depth. For the trapezoidal barrier, which applies to all bias conditions of interest in the present context, the electric field dependence is very small and is contained primarily in the exponential factor Eq. B.31. The electric field dependence does not become important until $|q\mathcal{E}x| > \phi_c - E$, when the tunneling barrier becomes triangular and changes in the electric field modulate the width of the barrier. In this case (field emission) an analysis similar to the above results in an exponential field dependence as in the Fowler Nordheim result, Eq. 3.1.

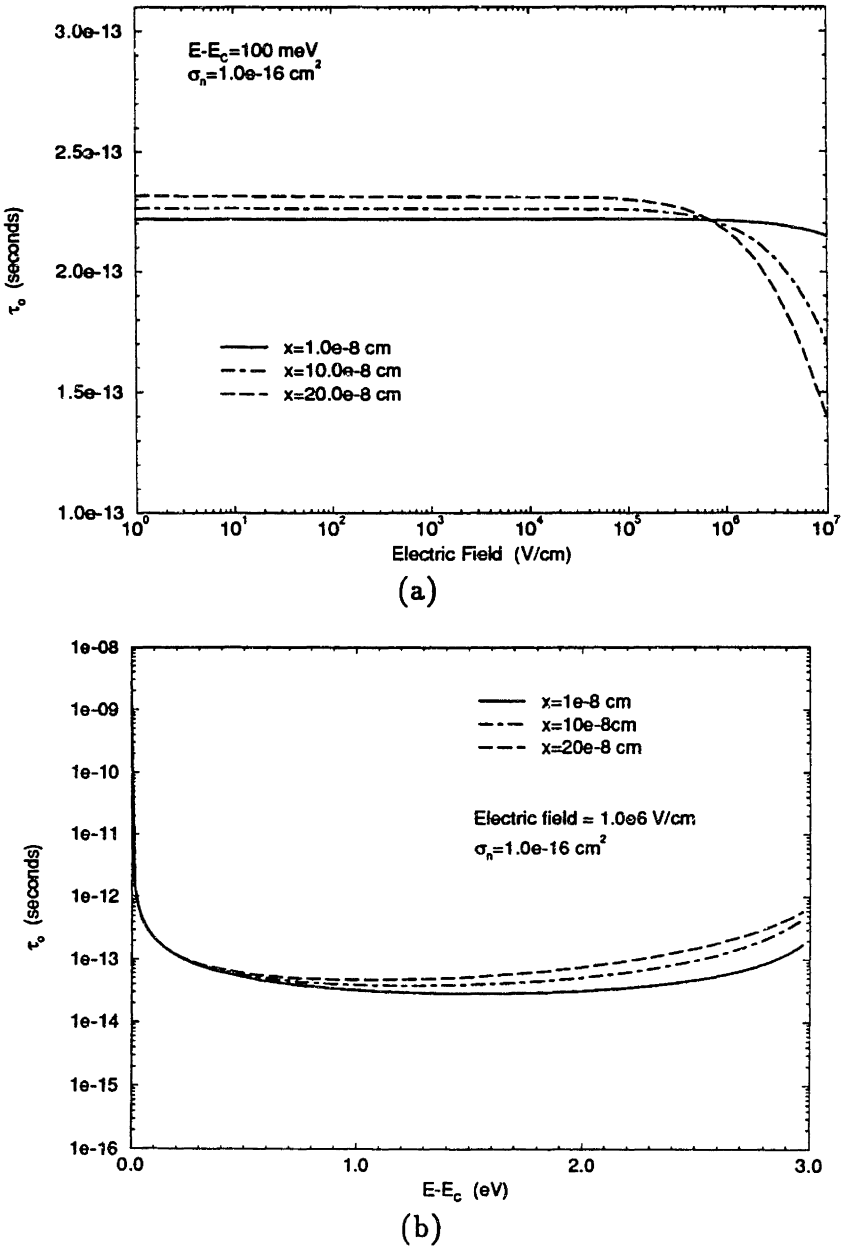


Figure B.3: Time constant prefactor (a) Electric field dependence (b) Energy dependence.

The time constant prefactor, Eq. B.32, is shown as a function of electric field for several values of x in Fig. B.3, assuming an oxide trap capture cross section of 10^{-16} cm^2 [130, 120, 234]. Since the field dependence is relatively weak, we shall neglect the electric field dependence and now turn to a three-dimensional model for the oxide trap potential.

B.1.2 Three-dimensional Oxide Trap Potential

In this section we assume a rectangular barrier and a spherically symmetric delta function for the oxide trap,

$$\begin{aligned} U_c(x) &= \phi_c \\ V_t(r) &= V_\delta \delta(r - r_o) \end{aligned} \tag{B.34}$$

where $r_o = (x_o, 0, 0)$.

The Schrödinger equation for ψ_l is, from Eq. B.5,

$$\left[-\frac{\hbar^2 \nabla^2}{2m^*} + \phi_c \theta(x) - E \right] \psi_l = 0 \tag{B.35}$$

The solutions for ψ_l are again of the form,

$$\psi_l(\vec{r}) = e^{i(k_y y + k_z z)} \chi_l(x) = e^{\vec{k}_{||} \cdot \vec{r}_{||}} \chi_l(x) \tag{B.36}$$

In the region $x < 0$, $U_c(x) = 0$ and $m^* = m_l$. Substituting Eq. B.36 into Eq. B.35 gives

$$\frac{d^2 \chi_l}{dx^2} + k_x^2 \chi_l = 0 \tag{B.37}$$

with solutions,

$$\chi_l(x) = A e^{i k_x x} + B e^{-i k_x x} \tag{B.38}$$

where

$$k_x{}^2 + \vec{k}_{||}{}^2 = k^2 = \frac{2 m_l E}{\hbar^2} \tag{B.39}$$

For $x > 0$ where $U_c(x) = \phi_c$ and $m^* = m_r$, we obtain,

$$\frac{d^2 \chi_l}{dx^2} - K_x^2 \chi_l = 0 \tag{B.40}$$

with the solution,

$$\chi_l(x) = C e^{-K_x x} \tag{B.41}$$

where

$$K_x^2 = K^2 + \vec{k}_{||}{}^2 \tag{B.42}$$

and

$$K^2 = \frac{2m_r}{\hbar^2}(\phi_c - E) \tag{B.43}$$

In Eq. B.41, the exponentially growing solution has been eliminated since the oxide is assumed to extend to infinity in the positive x direction and the wavefunction must be bounded.

The coefficients are determined by applying the boundary conditions necessary to ensure continuity of the wavefunctions and the current density in the effective mass approximation [235]:

$$\psi_l|_{x=0^-} = \psi_r|_{x=0^+} \tag{B.44}$$

$$\frac{1}{m_l}\frac{\partial \psi_l}{\partial x}\bigg|_{x=0^-} = \frac{1}{m_r}\frac{\partial \psi_r}{\partial x}\bigg|_{x=0^+} \tag{B.45}$$

Applying these equations to Eqs. B.38 and B.41, yields

$$\frac{A}{C} = \frac{1}{2}\left[1 + i\frac{m_l}{m_r}\frac{K_x}{k_x}\right] \tag{B.46}$$

$$\frac{B}{C} = \frac{1}{2}\left[1 - i\frac{m_l}{m_r}\frac{K_x}{k_x}\right] \tag{B.47}$$

The probability of the electron being in the insulator is small relative to its being in the semiconductor due to the exponential decay of the wavefunction, Eq. B.41. It is therefore a good approximation to normalize the electron wavefunction in the propagating region $x < 0$ only. This gives

$$\chi_l^* \chi_l \approx \frac{C^2}{4}\left\{2\left[1 + \left(\frac{m_l K_x}{m_r k_x}\right)^2\right] + \left(1 - i\frac{m_l K_x}{m_r k_x}\right)^2 e^{-2ik_x x} + \left(1 + i\frac{m_l K_x}{m_r k_x}\right)^2 e^{+2ik_x x}\right\} \tag{B.48}$$

Normalizing,

$$\int_{-L_x/2}^{+L_x/2} dz \int_{-L_y/2}^{+L_y/2} dy \int_{-L_x}^{0} dx \chi_l^* \chi_l = \frac{C^2 L_x L_y L_z}{2}\left[1 + \left(\frac{m_l K_x}{m_r k_x}\right)^2\right] = 1 \tag{B.49}$$

so that

$$C = \left[\frac{2{m_r}^2 {k_x}^2}{\Omega\left({m_r}^2 {k_x}^2 + {m_l}^2 K_x^2\right)}\right]^{1/2} \tag{B.50}$$

where the oscillating terms in Eq. B.48 contribute negligibly to the integral. Thus for $x > 0$,

$$\psi_l(\vec{r}) = \left[\frac{2{m_r}^2 {k_x}^2}{\Omega\left({m_r}^2 {k_x}^2 + {m_l}^2 K_x^2\right)}\right]^{1/2} e^{\vec{k}_{\|}\cdot\vec{r}_{\|}} e^{-K_x x} \tag{B.51}$$

Using H_r from Eq. B.6, Schrödinger's equation for ψ_r in spherical coordinates is

$$\left[-\frac{\hbar^2}{2m_i^*}\left(\frac{\partial^2}{\partial r'^2}+\frac{2}{r'}\frac{\partial}{\partial r'}+\frac{1}{r'^2}\mathcal{A}\right)+\phi_c+V_t(r')-E\right]\psi_r(r')=0 \tag{B.52}$$

where $r'=r-r_o$ and $\mathcal{A}$ is the angular operator

$$\mathcal{A}=\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)+\frac{1}{\sin\theta}\frac{\partial^2}{\partial\theta^2} \tag{B.53}$$

Since the trap potential, Eq. B.34, is a function of r' only, we assume a spherically symmetric $l=0$ (s) state $\psi_r(r',\theta,\phi)=\psi_r(r')$, so that Eq. B.52 can be written,

$$\left[-\frac{\hbar^2}{2m_i^*}\frac{1}{r'^2}\frac{d}{dr'}\left(r'^2\frac{d}{dr'}\right)+\phi_c+V_\delta\delta(r')-E\right]\psi_r(r')=0 \tag{B.54}$$

which has solutions of the form

$$\psi_r(r')=C\frac{e^{-Kr'}}{r'} \tag{B.55}$$

where, by direct substitution of Eq. B.55 into Eq. B.54,

$$K^2=\frac{2m_r}{\hbar^2}(\phi_c-E) \tag{B.56}$$

The constant C is determined from normalization,

$$\begin{aligned}\int_0^{2\pi}d\phi\int_0^{\pi}\sin\theta d\theta\int_0^{\infty}r'^2\psi_r{}^*(r')\psi_r(r')dr' &= 4\pi\int_0^{\infty}r^2\psi_r{}^*(r')\psi_r(r')dr' \\ &= \frac{2\pi C^2}{K}\end{aligned} \tag{B.57}$$

The trap wavefunction is therefore,

$$\psi_r(r')=\left(\frac{K}{2\pi}\right)^{1/2}\frac{e^{-Kr'}}{r'} \tag{B.58}$$

The matrix element is given by Eq. A.8,

$$M_{lr}=\int\psi_r{}^*(r')(H-E_l)\psi_l(r')d^3r' \tag{B.59}$$

where the exact Hamiltonian is,

$$H=-\frac{\hbar^2\nabla^2}{2m^*}+\phi_c\theta(x)+V_\delta\delta(r') \tag{B.60}$$

Since $H_l = H$ everywhere except at $r = r_o$, the value of the integral is zero everywhere except at $r' = 0$. Letting where ϵ is the radius of a small spherical surface surrounding the point $r' = 0$, Eq. B.59 can be rewritten,

$$M_{lr} = \lim_{\epsilon \to 0} \int_0^\epsilon \psi_r^*(r')(H - E)\psi_l(r')d^3r' = V_\delta \psi_r^*(0)\psi_l(0) \tag{B.61}$$

where all terms in the integrand except for the delta function are finite and hence go to zero as ϵ approaches zero.

To calculate M_{lr} in Eq. B.61, the strength of the delta function potential must be related to known quantities. Multiplying the Schrödinger equation Eq. B.54, by $4\pi r'^2$, integrating from $r' = 0$ to ϵ, and letting ϵ approach zero yields

$$\lim_{\epsilon \to 0} \frac{2\pi\hbar^2}{m_r} \left[\left(\frac{K}{2\pi} \right)^{1/2} (1 - \epsilon K) e^{-K\epsilon} - r'^2 \left. \frac{d\psi_r}{dr'} \right|_{r'=0} \right] = V_\delta \psi_r(0) \tag{B.62}$$

For a square potential well or radius a, the solution to Schrödinger's equation for $r' < a$ are $\psi_r(r') = \sin(kr')/r'$, so that

$$r'^2 \left. \frac{d\psi_r}{dr'} \right|_{r'=0} = (r'k \cos kr' - \sin kr')_{r'=0} = 0 \tag{B.63}$$

Since the delta function is just the limiting case of the square well as $a \to 0$, Eq. B.63 must continue to hold. The condition on the strength of the delta well potential, Eq. B.62 therefore becomes,

$$V_\delta \psi_r(r' = 0) = -\frac{\hbar^2}{m_r} \sqrt{2\pi K} \tag{B.64}$$

From Eqs. B.61 and B.64, the matrix element is

$$M_{lr} = \frac{\hbar^2}{m_r} \sqrt{2\pi K} \psi_l(r = r_o) \tag{B.65}$$

Substituting Eq. B.51 in Eq. B.65,

$$|M_{lr}|^2 = \frac{4\pi\hbar^4}{\Omega} \left(\frac{{k_x}^2 K}{{m_r}^2 {k_x}^2 + {m_l}^2 K_x^2} \right) e^{-2K_x x_o} \tag{B.66}$$

The tunneling rate can now be calculated by substituting the expression for the matrix element, Eq. B.66, into Eq. 3.17.

$$\begin{aligned} \omega(E, x) &= \frac{\Omega m_l}{\pi\hbar^3} \int \frac{|M_{lr}|^2}{k_x} \vec{k}_{||} d\vec{k}_{||} \\ &= \frac{4\hbar}{m_l} \int_0^{\vec{k}_{||_{max}}} \frac{k_x K}{{k_x}^2 \left(\frac{m_r}{m_l} \right)^2 + K_x^2} e^{-2K_x x} \vec{k}_{||} d\vec{k}_{||} \\ &= \frac{4\hbar}{m_l} \int_0^{\vec{k}_{||_{max}}} \frac{k_x K}{{k_x}^2 \left(\frac{{m_r}^2}{{m_l}^2} - 1 \right) + K^2 + k^2} e^{-2K_x x} \vec{k}_{||} d\vec{k}_{||} \end{aligned} \tag{B.67}$$

where Eqs. (B.39) and (B.42) have been used. Equation B.27 shows that the first term in the denominator can normally be neglected. Using Eq. B.39 and recognizing that $k_{max} = k$,

$$\omega(E,x) = \frac{4\hbar}{m_l}\frac{K}{K^2+k^2}\int_0^k (k^2 - \vec{k}_{||}^{\,2})^{1/2} e^{-2K_x x}\vec{k}_{||}d\vec{k}_{||} \tag{B.68}$$

Expanding K_x as in Eq. B.28, Eq. B.68 can be written,

$$\omega(E,x) = \frac{4\hbar}{m_l}\left(\frac{K}{K^2+k^2}\right)e^{-2Kx}\int_0^k (k^2 - \vec{k}_{||}^{\,2})^{1/2}\exp\left(-\frac{\vec{k}_{||}^{\,2}x}{K}\right)\vec{k}_{||}d\vec{k}_{||} \tag{B.69}$$

Carrying out the integration,

$$\omega(E,x) = \frac{2\hbar}{m_l x}\frac{kK^2}{K^2+k^2}\left\{1 - \left(\frac{K}{k^2x}\right)^{1/2}\mathcal{D}\left[\left(\frac{k^2x}{K}\right)^{1/2}\right]\right\}e^{-2Kx} \tag{B.70}$$

where

$$\mathcal{D}(x) = e^{-x^2}\int_0^x e^{t^2}dt \tag{B.71}$$

is a Dawson integral. This agrees with the result obtained by Lundstrom et al [234], except for the effective mass in the prefactor. The reason for the difference arises from the matching conditions use in the effective mass approximation, Eq.(B.45).

If the slowly varying factor $(k^2 - \vec{k}_{||}^{\,2})$ in the integrand of Eq. B.69 had been approximated by its value at $\vec{k}_{||} = 0$, due to the sharp peaking of the exponential, we would have obtained

$$\omega(E,x) \approx \frac{2\hbar}{m_l x}\frac{kK^2}{K^2+k^2}\left[1 - \exp\left(\frac{-k^2x}{K}\right)\right]e^{-2Kx} \tag{B.72}$$

which, for $\frac{k^2x}{K} > 1$ can be approximated by

$$\omega(E,x) \approx \frac{2\hbar}{m_l x}\frac{kK^2}{K^2+k^2}e^{-2Kx} \tag{B.73}$$

Apart from a constant prefactor, this agrees with the result obtained by Vuillaume et al. [72] using Oppenheimer's method.

Note that,in the case of a rectangular barrier, $K(x) = K$, equations B.72 and B.30 are identical if the cross section is taken to be

$$\sigma_n = 8\pi\left(\frac{m_r}{m_l}\right)^2\frac{k}{K(k^2+K^2)} \approx 8\pi\left(\frac{m_r}{m_l}\right)^2\frac{k}{K^3} \tag{B.74}$$

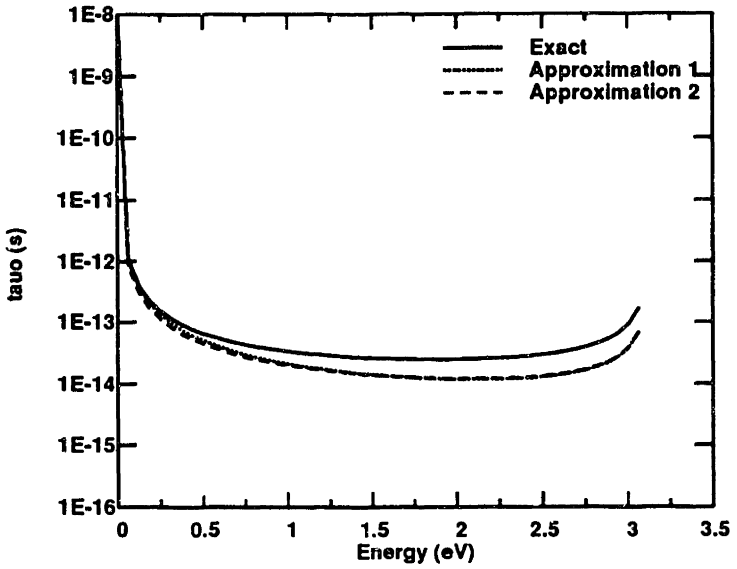


Figure B.4: Time constant versus energy above the conduction band for a trap at a depth of 50Å from the interface, using Eq. B.69 (solid line), Eq. B.71 (dotted line) and Eq. B.74 (dashed line)

For small x the exponentials in the Dawson integral, Eq. (B.71) can be expanded to give

$$\omega(E,x) \approx \frac{4\hbar}{3m_l}\frac{k^3K}{K^2+k^2}\left(1-\frac{2}{5}\frac{k^2x}{K}\right)e^{-2Kx} \tag{B.75}$$

The dominant dependence of the trapping time constant $\tau = 1/\omega$ is due to the exponential in Eq.(B.75), however the prefactor is also seen to depend weakly on energy and depth as investigated in Chapter 3. The energy dependence of the time constant prefactor τ_o is shown in Fig. B.4. The time constant changes rapidly near the conduction band edge. The approximations, B.72 and B.75 are in satisfactory agreement with the exact expression, Eq. B.70 in the energy region of interest just above the conduction band edge.

B.2 Interface State to Oxide Trap Tunneling

We now turn to the problem of tunneling between interface state and oxide traps. The general features of the problem can best be understood by first considering the problem of a particle in a double well potential as shown in Fig. B.5 The double well can be solved exactly [236], but since the coupling between the wells is weak, an approximate solution can be found by constructing a solution from a linear combinations of the solutions for the isolated wells [237, 238]. A consideration of this approach is useful

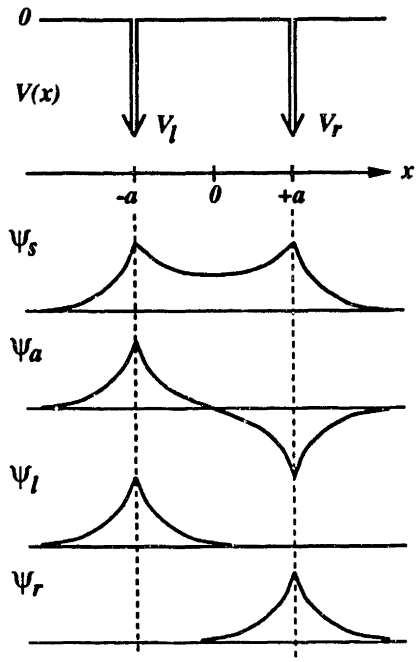


Figure B.5: Double-well potential

in order to establish the connection between Bardeen's method and the conventional description of the two-state problem in quantum mechanics [239]. Bardeen's method will then be used in Secs. B.2.1 and B.2.2 to evaluate the tunneling time constant for this process.

The double well has been treated by a number of authors [238, 240, 241, 242, 243, 244]. If the barrier were inpenetrable, the solutions to Schrödinger's equation in each well are oscillating inside and exponentially decaying outside. In the case where the wells are symmetrical, $V_r(x) = V_l(-x) = V(x)$, the energy levels are doubly degenerate, corresponding to motion of the particle in one well or the other for all time.

$$\begin{aligned} H_l\psi_l^n &= E_l\psi_l^n \\ H_r\psi_r^m &= E_r\psi_r^m \end{aligned} \tag{B.76}$$

where the integers m and n designate the set of solutions to Schrödinger's equation in each well. It will be assumed, for simplicity, that tunneling occurs directly into the ground state $m = n = 0$. The wavefunctions ψ_l and ψ_r are assumed individually orthonormal but are not orthogonal to each other because they are eigenfunctions of different Hamiltonians. The total Hamiltonian for the system is

$$H = -\frac{\hbar^2}{2m}\nabla^2 + V_l(x) + V_r(x)$$

$$\begin{aligned} &= H_l + V_r(x) \\ &= H_r + V_l(x) \end{aligned} \tag{B.77}$$

When the wells are coupled, ψ_l and ψ_r are no longer stationary states. An approximate trial solution to the Schrödinger equation can be written as a linear combination of the solutions for the uncoupled wells [237, 238, 245],

$$|\psi\rangle = C_1|\psi_l\rangle + C_2|\psi_r\rangle \tag{B.78}$$

The stationary states satisfy the time independent Schrödinger equation,

$$H|\psi\rangle = E|\psi\rangle \tag{B.79}$$

Multiplying Eq. B.79 on the left by $\langle\psi_l|$ and $\langle\psi_r|$ yields the eigenvalue equation

$$\begin{vmatrix} H_{11} - E & H_{12} - E\Delta \\ H_{21} - E\Delta & H_{22} - E \end{vmatrix} \tag{B.80}$$

where $\langle\psi_l|\psi_l\rangle = \langle\psi_r|\psi_r\rangle = 1$ and,

$$\Delta = \langle\psi_l|\psi_r\rangle = \langle\psi_r|\psi_l\rangle \tag{B.81}$$

$$H_{11} = E_l + \langle\psi_l|V_r|\psi_l\rangle \tag{B.82}$$

$$H_{12} = E_r\Delta + \langle\psi_l|V_l|\psi_r\rangle \tag{B.83}$$

$$H_{21} = E_l\Delta + \langle\psi_r|V_r|\psi_l\rangle \tag{B.84}$$

$$H_{22} = E_r\langle\psi_r|V_l|\psi_r\rangle \tag{B.85}$$

In the case of a symmetric well, $E_l = E_r = E_o$, $H_{11} = H_{22}$ and $H_{12} = H_{21}$ and the eigenvalues of Eq. B.80 are

$$E = \frac{H_{11} \pm H_{12}}{1 \pm \Delta} \tag{B.86}$$

For $x < 0$, $\psi_r \ll \psi_l$, while the opposite is true for $x > 0$, thus $\Delta \ll 1$ everywhere and the wavefunctions are "almost orthogonal". The eigenvalues are then

$$E_s = E_o + \langle\psi_l|V_r|\psi_l\rangle - \langle\psi_l|V_l|\psi_r\rangle \tag{B.87}$$

$$E_a = E_o + \langle\psi_l|V_r|\psi_l\rangle + \langle\psi_l|V_l|\psi_r\rangle \tag{B.88}$$

The corresponding eigenvectors are easily found to be

$$\psi_s \approx \frac{1}{\sqrt{2}}(\psi_l + \psi_r) \tag{B.89}$$

$$\psi_a \approx \frac{1}{\sqrt{2}}(\psi_l - \psi_r) \tag{B.90}$$

Thus, the coupling of the two wells splits each of the uncoupled energy levels into two, separated by an amount $E_a - E_s = 2\langle\psi_l|V_l|\psi_r\rangle$. The stationary states consist of a symmetric wavefunction ψ_s corresponding to the lower energy level and an antisymmetric wavefunction ψ_a corresponding to the higher energy level.

When an interface state captures an electron from the conduction band, the electron wavefunction is initially localized in one of the wells. Thus, consider preparing a wavefunction, which is localized, as much as possible, in the left well at $t = 0$,

$$\psi_l(x) = \frac{1}{\sqrt{2}}(\psi_s + \psi_a) \tag{B.91}$$

Since ψ_s and ψ_a have different frequencies, this is not a stationary state, but will change with time as

$$\psi_l(x,t) = \frac{1}{\sqrt{2}}\left[\psi_s(x)e^{-iE_st/\hbar} + \psi_a(x)e^{-iE_at/\hbar}\right] \tag{B.92}$$

$$|\psi_l(x,t)|^2 = \frac{1}{2}\left[\psi_s^2 + \psi_a^2 + (\psi_s^2 - \psi_a^2)\cos\left(\frac{E_a - E_s}{\hbar}t\right)\right] \tag{B.93}$$

The interference between ψ_s and ψ_r causes the wavefunction to oscillate back and forth between the two wells with frequency proportional to the splitting of the energy levels $\omega = (E_a - E_s)/\hbar$. In the two-step tunneling process, the oscillation does not occur because the thermal process occurs much faster than the tunneling process and, once the electron tunnels out of the interface state, another electron is captured from the conduction band. Similarly, when an electron tunnels from an oxide trap to an interface state, the electron is "instantaneously" swept away be thermal emission into the conduction band. The wavefunction, which is localized on the left well at $t = 0$ will have tunneled through the barrier to the right well after a time

$$\begin{aligned} t &= \frac{\pi\hbar}{E_a - E_s} \\ &= \frac{\pi\hbar}{\langle\psi_l|V_l|\psi_r\rangle} \end{aligned} \tag{B.94}$$

The matrix element in the denominator of Eq. B.94 is the same as that which appears in Bardeen's effective tunneling matrix element Eq. A.8, since from Eq. B.77,

$$\langle\psi_l|V_l|\psi_r\rangle = \langle\psi_l|H - H_r|\psi_r\rangle = \langle\psi_r|H - H_l|\psi_l\rangle = M_{lr} \tag{B.95}$$

This is not surprising since the same expansion, Eq. B.78 was used in both analyses, based on the assumption that the uncoupled wavefunctions are nearly orthogonal. Thus, the tunneling time depends exponentially on the spatial separation of the two traps.

If the wells are asymmetrical then their uncoupled energy levels will differ. A transistion from a state localized in one well to a state localized in the other well then

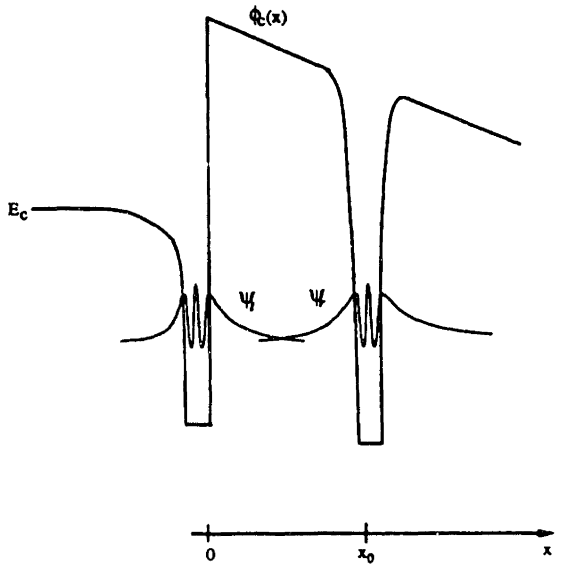


Figure B.6: Band diagram showing interface and oxide trap potential wells.

must be stimulated by some time dependent perturbation with frequency $E_r - E_l$ in order to conserve energy [239]. Phonon-assisted tunneling between localized states have been treated by many authors [241]. This process leads to thermally activated "hopping" processes which can be modeled as [115]

$$T \propto \nu e^{-E_a/kT} T \tag{B.96}$$

where T is the WKB transmission probability through the barrier. In accordance with the discussion in Sec. 3.3, we neglect all but the elastic tunneling process described above. Since the interface states are continuously distributed within the bandgap, an oxide trap which is lowered to within the energy range of the silicon bandgap by an applied electric field has a high probability of lining up with an interface state at the same energy.

In the case of tunneling from an interface state to a trap in the insulator, the potential is not symmetric due to the discontinuity in band structure at the interface and the energy eigenfunctions do not have definite parity. The tunneling time in Eq. B.94 is related to, but not identical to, the tunneling time constant as defined in Sec. 3.3. We therefore turn to Bardeen's method for the calculation of the tunneling time constant.

Figure B.6 shows the energy band diagram for the Si-SiO_2 system, with traps at the interface and in the insulator. The exact Schrödinger equation is,

$$\left[-\frac{\hbar^2\nabla^2}{2m^*} + U_c(x)\theta(x) + V_{it}(r) + V_{ot}(r - r_o) - E\right]\psi = 0$$

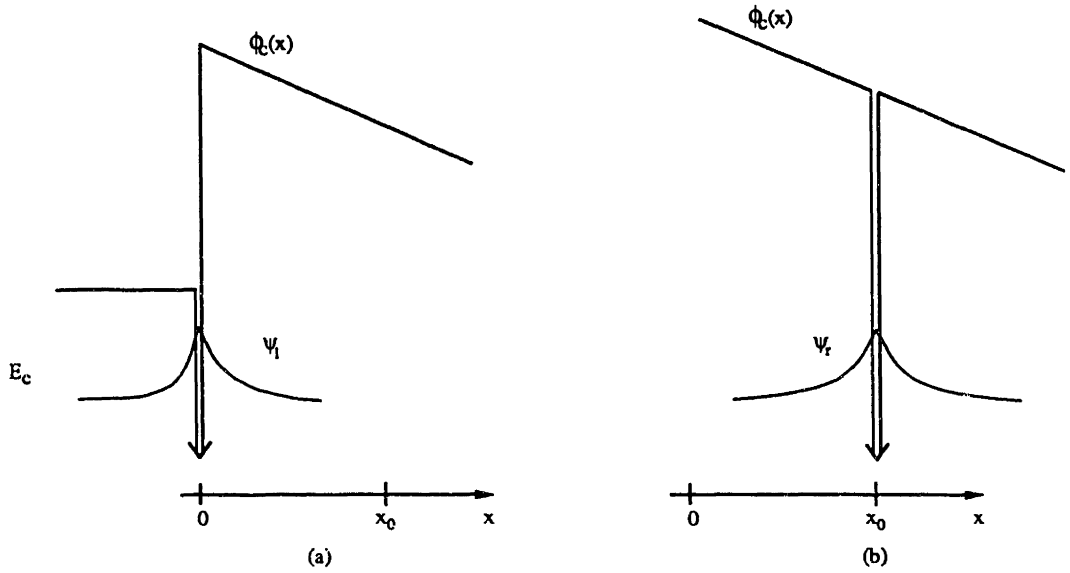


Figure B.7: Separation of the exact Hamiltonian into two partial Hamiltonians for (a) left side wavefunction (interface state) (b) right side wavefunction (oxide trap)

where V_{it} and V_{ot} are the potentials of the interface and oxide traps, located at $r = (0,0,0)$ and $r = (x_o,0,0)$ respectively. Figure B.7 shows the separation of the total Hamiltonian into two uncoupled systems for the two wells. The left side wavefunction is a bound state of the interface potential well, as shown in Fig. B.7(a), with Hamiltonian

$$H_l = -\frac{\hbar^2\nabla^2}{2m^*} + U_c(x)\theta(x) + V_{it}(r) \tag{B.97}$$

The right side wavefunction is a bound state of an oxide trap in an insulator of infinite thickness, as shown in Fig. B.7(b), with Hamiltonian

$$H_r = -\frac{\hbar^2\nabla^2}{2m^*} + U_c(x) + V_{ot}(r) \tag{B.98}$$

In the following two sections, Bardeen's method is used to calculate the tunneling time constant, first assuming a one-dimensional potential for the oxide trap and then using a three-dimensional delta function.

B.2.1 One–Dimensional Oxide Trap Potential

We begin, as in Section B.1.1, by assuming a trapezoidal barrier and a one dimensional delta function potential for both the interface and oxide traps,

$$U_c(x) = \phi_c + q\mathcal{E}x \tag{B.99}$$

$$V_{it}(r) = V_\delta \delta(x) \tag{B.100}$$
$$V_{ot}(r) = V_\delta \delta(x - x_o) \tag{B.101}$$

The right side wavefunction ψ_r is again given by Eqs. B.16 and B.16,

$$\psi_r(x) = \begin{cases} \frac{C_r}{\sqrt{K(x)}} \exp\left[-\int_x^{x_o} K_x(x')dx'\right] & x < x_o \\ \frac{C_r}{\sqrt{K(x)}} \exp\left[-\int_{x_o}^{x} K_x(x')dx'\right] & x > x_o \end{cases} \tag{B.102}$$

with $C_r = \frac{K(x_o)}{A^{1/2}}$. The left side wavefunction is

$$\psi_l(x) = \begin{cases} \frac{C_{l1}}{\sqrt{\kappa(x)}} \exp\left[-\int_x^0 \kappa(x')dx'\right] & x < 0 \\ \frac{C_{l2}}{\sqrt{K(x)}} \exp\left[-\int_0^x K(x')dx'\right] & x > 0 \end{cases} \tag{B.103}$$

where $\vec{k}_{||}$ has been set equal to zero in Eq. B.9 because we are now dealing with bound states rather than traveling waves at the interface. Thus, from Eqs. B.97–B.103,

$$\kappa(x)^2 = \frac{2m_l}{\hbar^2}|E| \tag{B.104}$$
$$K(x)^2 = \frac{2m_r}{\hbar^2}[U_c(x) - E] \tag{B.105}$$

where $E < 0$ since energies are measured upward from E_c. C_{l1} and C_{l2} are obtained by matching the wavefunction ψ_l across the interface, $x = 0$ and normalizing to unity,

$$C_{l2} = \left(\frac{K(0)}{\kappa(0)}\right)^{1/2} C_{l1} \tag{B.106}$$

Since the wavefunction decays exponentially away from $x = 0$, $K(x) \approx K(0)$ and $\kappa(x) \approx \kappa(0)$. Normalizing,

$$\int_{-L_x/2}^{L_x/2} dz \int_{-L_y/2}^{L_y/2} dy \int_{-\infty}^{+\infty} \psi_l^* \psi_l dx = A\frac{C_{l1}^2}{2\kappa(0)}\left[\frac{1}{\kappa(0)} + \frac{1}{K(0)}\right] = 1 \tag{B.107}$$

so that

$$C_{l1} = (2\kappa(0)\kappa'/A)^{1/2} \tag{B.108}$$
$$C_{l2} = (2K(0)\kappa'/A)^{1/2} \tag{B.109}$$

where

$$\kappa' = \frac{\kappa(0)K(0)}{\kappa(0) + K(0)} \tag{B.110}$$

Substituting the wavefunctions, Eqs. B.103 and B.102 into Eq. A.16 , the matrix element is

$$M_{lr} = \frac{\hbar^2 A}{m_r} C_{l2} C_r \exp\left[-\int_0^x K(x')dx'\right] \quad \text{(B.111)}$$

$$= \frac{\hbar^2}{m_r}(2K(0)\kappa')^{1/2} K(x) \exp\left[-\int_0^x K(x')dx'\right] \quad \text{(B.112)}$$

Multiplying Eq. 3.42 by σ_n/A as in Sec. B.1.1 to account for the fact that the one-dimensional oxide trap covers the entire plane

$$\omega(E,x) = \frac{2\pi}{\hbar}\sigma_n D_{it}(E)|M_{lr}|^2 \quad \text{(B.113)}$$

$$= \frac{4\pi\hbar^3\sigma_n D_{it}}{{m_r}^2}\frac{\kappa(0)K(0)^2K(x)^2}{\kappa(0)+K(0)} \exp\left[-2\int_0^x K(x')dx'\right] \quad \text{(B.114)}$$

B.2.2 Three Dimensional Oxide Trap Potential

A better model for the oxide trap is a 3-dimensional delta function well as in Sec. B.1.2. As in that section, we assume a rectangular tunneling barrier,

$$U_c(x) = \phi_c \quad \text{(B.115)}$$

$$V_{it}(r) = V_\delta \delta(x) \quad \text{(B.116)}$$

$$V_{ot}(r) = V_\delta \delta(r - r_o) \quad \text{(B.117)}$$

For the interface state we continue to assume a one-dimensional well covering the whole interface plane at $x = 0$. In addition to circumventing the need for solving a spherical problem across a discontinuity in band structure, this model, which treats the interface states as an averaged continuous distribution in space, is appropriate since the actual location of the interface states with respect to the oxide traps is unknown.

The oxide trap wavefunction is, as before by Eq. B.102

$$\psi_r(r) = \left(\frac{K}{2\pi}\right)^{1/2} \frac{e^{-K|r-r_o|}}{|r-r_o|} \quad \text{(B.118)}$$

where $r_o = (x, 0, 0)$. The interface state wavefunction is

$$\psi_l(x) = \begin{cases} Ce^{\kappa x} & x < 0 \\ Ce^{-\kappa x} & x > 0 \end{cases} \quad \text{(B.119)}$$

where K and κ are defined in Eqs. B.104 and B.105. The constants are found by normalization

$$C^2 \int_{-L_z/2}^{L_z/2} dz \int_{-L_y/2}^{L_y/2} dy \left[\int_{\infty}^{0} e^{\kappa x} dx + \int_0^{\infty} e^{-2Kx} dx\right] = 1 \quad \text{(B.120)}$$

$$C = \left[\frac{2\kappa K}{A(\kappa + K)}\right]^{1/2} \tag{B.121}$$

From Eq. B.65, the matrix element is

$$M_{lr} = \frac{\hbar^2}{m_\tau}\sqrt{2\pi K}\psi_l(r_o) \tag{B.122}$$

$$= \frac{2\hbar^2}{m_\tau}\left[\frac{\pi\kappa K^2}{A(\kappa + K)}\right]^{1/2} e^{-Kx} \tag{B.123}$$

From Eq. 3.42 the transition rate per unit time is

$$\omega(E, x) = \frac{2\pi}{\hbar} D_{it}(E) {M_{lr}}^2 A \tag{B.124}$$

$$= \frac{8\pi^2\hbar^3 D_{it}(E)}{{m_\tau}^2}\left(\frac{\kappa K^2}{\kappa + K}\right) e^{-2Kx} \tag{B.125}$$

Assuming $m_\tau = 0.42m_o$ and $m_l = 0.1m_o$, the time constant prefactor is found to be $\tau_o(E) = 5.33 \times 10^{-3}/D_{it}(E)$ which, for $D_{it} = 10^9$ cm^{-2} eV^{-1}, gives $\tau_o = 5 \times 10^{-12}$s.

The one-dimensional and three-dimensional results, Eqs. B.114 and B.125, will be equal in the case of a rectangular barrier if $\sigma_n = 2\pi/K^2$. Thus, $1/K$ can be interpreted as the capture radius of the oxide trap. This is only a consequence of the delta function model assumed for the oxide trap and should not be literally interpreted as a physical cross section. Nevertheless, with the values given above, this implies an oxide trap capture cross section of approximately 8.0×10^{-14} cm^2 which is the correct order of magnitude. For modeling purposes, it is more useful to express the time constant in terms of σ_n (Eq. B.114) in order to provide an adjustable fitting parameter.

Appendix C

Hole Tunneling and Lattice Relaxation

In this appendix the analysis of Secs. 3.3 and 3.4.3 is extended to include elastic tunneling transitions of holes from the valence band into oxide traps, followed by relaxation of the trap to its ground state energy. From Sec. 3.3, the rate of change of the oxide trap occupancy function $f_r(E_r)$ due to electron tunneling is described by the differential equation [Eq. 3.10]

$$\left.\frac{\partial f_r(E_r)}{\partial t}\right|_{electron} = \frac{f_l(E_r) - f_r(E_r)}{\tau_n} \tag{C.1}$$

where the subscript n on the time constant denotes electron tunneling, to be distinguished from hole tunneling as discussed below. For tunneling from the conduction band, the time constant is given by Eq. 3.9, integrated over all conduction band energies,

$$\frac{1}{\tau_n} = \{1 + \exp[(E_r - F_l)/kT\} \sum_{\vec{k}_{||}} \int_{E_c}^{\infty} W_{lr}(E_l, E_r, \vec{k}_{||}) \rho_l(E_l) f_l(E_l) dE_l$$

With the transition rate W_{lr} given by Eq. 3.22, the time constant is [Eqs. 3.27, 3.29]

$$\tau_n = \tau_{no} e^{\lambda_c(x)}$$

where $\lambda_c(x)$ is the WKB integral evaluated at the conduction band edge,

$$\lambda_c(x) = \frac{4}{3} \left(\frac{2M_c^*}{\hbar^2}\right)^{1/2} \frac{(\phi_c + q\mathcal{E}x)^{3/2} - \phi_c^{3/2}}{q\mathcal{E}}$$

and the prefactor is

$$\tau_{no} = \frac{e^{\Delta E_{Bc}/kT}}{\sigma_n \overline{v}(n_s + n_1)}$$

where the barrier height for electron tunneling is ΔE_{Bc}.

It is also possible for a hole to tunnel elastically into an excited state of an oxide trap, followed by a relaxation to the ground state. In the case of hole traps, relaxation means that the trap, after capturing a hole (or emitting an electron), rises relative to the valence band edge, since hole energies are measured downward. This is consistent with the case of electron traps, where the presence of an electron caused a decrease in energy.

For the trapping of holes, it is convenient to define

$$f_l'(E_l) = 1 - f_l(E_l) = \frac{1}{1 + e^{-(E_l - F_l)/kT}} \tag{C.2}$$

$$f_r'(E_r) = 1 - f_r(E_r) = \frac{1}{1 + e^{-(E_r - F_r)/kT}} \tag{C.3}$$

The rate at which holes tunnel from valence band states at energy E_l into oxide traps at energy E_r is then

$$r_{lr}'(E_l, E_r, \vec{k}_{||}) = W_{lr}' f_l'(E_l) \rho_l(E_l)[1 - f_r'(E_r)]\rho_r(E_r)$$

while the rate at which holes tunnel out of traps is

$$r_{rl}' = W_{rl}' f_r'(E_r)\rho_r(E_r)[1 - f_l'(E_l)]\rho_l(E_l)$$

In equilibrium, when $f_l' = f_r' = f$ and $F_l = F_r = E_F$, these rates must self balance, $r_{lr}' = r_{rl}'$. This implies that

$$W_{lr}' = W_{rl}' e^{(E_r - E_l)/kT} \tag{C.4}$$

The total rate at which holes flow into oxide traps at energy E_r is then

$$\begin{aligned} r'(E_r) &= \sum_{\vec{k}_{||}} \int_{-\infty}^{E_v} (r_{lr}' - r_{rl}')dE_l \\ &= \sum_{\vec{k}_{||}} \int_{-\infty}^{E_v} W_{lr}'(E_l, E_r, \vec{k}_{||})\rho_l(E_l)\rho_r(E_r) \\ &\qquad \times \left\{ f_l'(E_l)[1 - f_r'(E_r)] - e^{(E_l - E_r)/kT} f_r'(E_r)[1 - f_l'(E_l)] \right\} dE_l \\ &= \frac{\partial}{\partial t}[f_r'(E_r)\rho_r(E_r)] \end{aligned} \tag{C.5}$$

using

$$1 - f_l'(E_l) = e^{-(E_l - F_l)/kT} f_l'(E_l)$$

Equation C.5 can be written

$$\left.\frac{\partial f_r'(E_r)}{\partial t}\right|_{hole} = \frac{f_l'(E_r) - f_r'(E_r)}{\tau_p} \tag{C.6}$$

where

$$\frac{1}{\tau_p} = \{1 + \exp[(F_l - E_r)/kT]\} \sum_{\vec{k}_{||}} \int_{-\infty}^{E_v} W'_{lr}(E_l, E_r, \vec{k}_{||}) \rho_l(E_l) f'_l(E_l) dE_l \qquad \text{(C.7)}$$

The transition rate W'_{rl} for holes from the trap to the valence band is given by the product of a Boltzmann factor for probability of excitation of the hole trap to the valence band, and a tunneling rate as given by Bardeen's matrix element,

$$W'_{rl}(E_r, E_l, \vec{k}_{||}) = \left\{ \frac{1}{kT} e^{-(E_r - E_l + \Delta E_{Bv})/kT} \right\} \left\{ \frac{2\pi}{\hbar} \left| M'_{lr}(E_l, \vec{k}_{||}) \right|^2 \right\}$$

where E_{Bv} is the barrier height for transitions from the valence band. Using the detailed balance condition [Eq. C.4],

$$W'_{lr} = \frac{2\pi}{\hbar kT} e^{-\Delta E_{Bv}/kT} |M'_{lr}|^2$$

Substituting this value for W'_{lr} into Eq. C.7, the time constant for hole tunneling is

$$\frac{1}{\tau_p} = \frac{2\pi}{\hbar kT} \left[1 + e^{(F_l - E_r)/kT}\right] e^{-\Delta E_{Bv}/kT} \sum_{\vec{k}_{||}} \int_{-\infty}^{E_v} dE_l \, |M'_{lr}|^2 \, \rho_l(E_l) f'_l(E_l) dE_l$$

The integration can be carried out in a manner analogous to Sec. 3.4.3 Under the assumption that M'_{lr} varies slowly with E_l and is independent of $\vec{k}_{||}$, we obtain

$$\begin{aligned} \frac{1}{\tau_p} &= \frac{2\pi}{\hbar kT} e^{-\Delta E_{Bv}/kT} \left[1 + e^{(F_l - E_r)/kT}\right] |M'_{lr}|^2 \int_{-\infty}^{E_v} g_v(E_l)[1 - f_l(E_l)] dE_l \\ &= \frac{2\pi}{\hbar kT} e^{-\Delta E_{Bv}/kT} (p_s + p_1) |M'_{lr}|^2 \end{aligned} \qquad \text{(C.8)}$$

where $g_v(E)$ is the three-dimensional density of valence band states,

$$g_v(E) = \frac{m_v^* \sqrt{2m_v^*(E_v - E)}}{\pi^2 \hbar^3}$$

and

$$\begin{aligned} p_s &= \int_{-\infty}^{E_v} g(E)[1 - f(E)] dE \\ p_1 &= p_s e^{-(E_r - F_l)/kT} \end{aligned}$$

The evaluation of the matrix element leads, as usual, to a WKB integral due to the overlap of the wavefunctions [Appendix B.1.1]. The remaining terms are lumped into a capture cross section as described in Sec. 3.4.3. Evaluating the WKB integral at

the valence band edge where the tunneling transition occurs, the hole time constant is

$$\tau_p = \tau_{po} e^{\lambda_v(x)}$$

where

$$\lambda_v(x) = \exp\left\{\frac{4}{3}\left(\frac{2M_v^*}{\hbar^2}\right)^{1/2} \frac{\phi_v^{3/2} - (\phi_v - q\mathcal{E}x)^{3/2}}{q\mathcal{E}}\right\}$$

and the prefactor of the exponential is

$$\tau_{po} = \frac{e^{\Delta E_{Bv}/kT}}{\sigma_p \overline{v}[p_s + p_1]}$$

A more rigorous calculation, using the one-dimensional density of valence band states and carrying out the summation over $\vec{k}_{||}$, leads to a result of the same general form, as shown in Sec. 3.4.3 for electron tunneling. The capture cross section depends on the form assumed for the oxide trap potential and can be estimated by the methods described in Sec. 3.4.3.

Combining the electron and hole tunneling contributions, Eqs. C.1 and C.6, the rate of change of the occupancy of oxide traps is

$$\frac{\partial(f_t \rho_t)}{\partial t} = \left.\frac{\partial(f_t \rho_t)}{\partial t}\right|_{electron} - \left.\frac{\partial(f_t' \rho_t)}{\partial t}\right|_{hole}$$

or, using Eqs. C.2 and C.3

$$\frac{\partial f_t(E_t)}{\partial t} = \frac{f_l(E_t) - f_t(E_t)}{\tau_{ot/lr}}$$

where the time constant for lattice relaxation, including tunneling transitions from both silicon bands, is

$$\begin{aligned} \frac{1}{\tau_{ot/lr}} &= \frac{1}{\tau_p} + \frac{1}{\tau_n} \\ &= \sigma_n e^{-\Delta E_{Bc}/kT} \overline{v}(n_s + n_1) e^{-\lambda_c(x)} + \sigma_p e^{-\Delta E_{Bv}/kT} \overline{v}(p_s + p_1) e^{-\lambda_v(x)} \end{aligned} \tag{C.9}$$

Note that this result is formally identical to a Shockley-Read-Hall analysis [28], with space and temperature-dependent capture cross sections:

$$\begin{aligned} \sigma_n(x,T) &= \sigma_n e^{-\Delta E_{Bc}/kT}\, e^{-\lambda_c(x)} \\ \sigma_p(x,T) &= \sigma_p e^{-\Delta E_{Bv}/kT}\, e^{-\lambda_v(x)} \end{aligned}$$

It is convenient to remove the dominant spatial dependence from Eq. C.9 and write the time constant in the standard form,

$$\tau_{ot/lr} = \tau_o e^{\lambda_c(x)} \tag{C.10}$$

where

$$\tau_o = \sigma_n e^{-\Delta E_{Bc}/kT} \overline{v}(n_s + n_1) + \sigma_p e^{-\Delta E_{Bv}/kT} \overline{v}(p_s + p_1) e^{\lambda_v(x)/\lambda_c(x)} \tag{C.11}$$

Appendix D

Solution of Coupled Rate Equations

In this appendix we solve the coupled rate equations Eqs. 3.62 and 3.63 in the case of a density of oxide traps N_{ot} cm^{-2} eV^{-1} located at depth x into the oxide. The equations for the time rates of change of the occupied oxide traps and interface states (cm^{-2} eV^{-1}) become

$$\frac{\partial n_{ot}(x,t)}{\partial t} = a(x)n_{it}(t) - b(x)n_{ot}(x,t) \tag{D.1}$$

$$\frac{\partial n_{it}(t)}{\partial t} = cn_{it}(t) + d - \frac{\partial n_{ot}(x,t)}{\partial t} \tag{D.2}$$

where

$$a(x) = T(x)N_{ot}(x) \tag{D.3}$$

$$b(x) = T(x)D_{it} = 1/\tau_{ot/it} \tag{D.4}$$

$$c = -[c_{ns}(n_s + n_1) + c_{ps}(p_s + p_1)] = -1/\tau_{it} \tag{D.5}$$

$$d = D_{it}[c_{ns}n_s + c_{ps}p_1] \tag{D.6}$$

In order to eliminate the constant term d, it is convenient to make the following substitutions,

$$n'_{it} = n_{it} + \frac{d}{c} \tag{D.7}$$

$$n'_{ot} = n_{ot} + \frac{ad}{bc} \tag{D.8}$$

The equations D.1 and D.2 are then transformed into

$$\frac{\partial n'_{ot}}{\partial t} = a(x)n'_{it} - b(x)n'_{ot}(x,t)$$

$$\frac{\partial n'_{it}}{\partial t} = cn'_{it}(t) - \frac{\partial n'_{ot}(xt)}{\partial t}$$

Assuming solutions of the form,

$$n'_{ot}(x,t) = Ae^{st} \tag{D.9}$$
$$n'_{it}(t) = Be^{st} \tag{D.10}$$

Substituting into the above equations A and B satisfy

$$\begin{bmatrix} s & s-c \\ s+b & -a \end{bmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = 0 \tag{D.11}$$

The eigenvalues s_1 and s_2 satisfy the characteristic equation

$$s^2 + s(a+b-c) - bc = 0 \tag{D.12}$$

The general solution to the original equations Eqs. D.1 and D.2 can be written, using Eqs. D.7–D.11, as

$$n_{it}(t) = A_1 e^{s_1 t} + A_2 e^{s_2 t} - \frac{d}{c}$$
$$n_{ot}(x,t) = \frac{aA_1}{s_1+b} e^{s_1 t} + \frac{aA_2}{s_2+b} e^{s_2 t} - \frac{ad}{bc}$$

where s_1 and s_2 are the roots of Eq. D.12. The constants A_1 and A_2 are determined by the initial conditions

$$A_1 = \frac{s_1+b}{s_1-s_2}\left[n_{it}(0) - \frac{ds_2}{bc} - \frac{s_2+b}{a} n_{ot}(x,0)\right]$$
$$A_2 = \frac{s_2+b}{s_1-s_2}\left[\frac{s_1+b}{a} n_{ot}(x,0) + \frac{ds_1}{bc} - n_{it}(0)\right]$$

Rewriting in terms of the time constants $\tau = -1/s$ and using the definitions D.3–D.6, the solutions are

$$f_{it}(E,t) = f(E) + [f_o(E) - f(E)]\left[\left(\frac{\tau_1}{\tau_1-\tau_2}\right) e^{-t/\tau_1} - \left(\frac{\tau_2}{\tau_1-\tau_2}\right) e^{-t/\tau_2}\right]$$
$$f_{ot}(E,x,t) = f(E) + [f_o(E) - f(E)]\left[\left(\frac{\tau_1-\tau_{ot/it}}{\tau_1-\tau_2}\right) e^{-t/\tau_1} - \left(\frac{\tau_2-\tau_{ot/it}}{\tau_1-\tau_2}\right) e^{-t/\tau_2}\right]$$

where τ_1 and τ_2 satisfy Eq. D.12,

$$\tau = \frac{1}{2}(\tau_{it} + \tau_{ot/it} + \tau_{it}\tau_{ot/it} T N_{ot})\left\{1 \pm \left[1 - \frac{4\tau_{it}\tau_{ot/it}}{(\tau_{it} + \tau_{ot/it} + \tau_{it}\tau_{ot/it} T N_{ot})^2}\right]^{1/2}\right\}$$

with $f(E)$ and $f_o(E)$ given by Eqs. 3.75 and 3.76.

Appendix E

Band bending Calculations

The model of Chap. 3 requires the Fermi level, or equivalently, the band bending at the surface as a function of the applied gate voltage. The gate voltage is related to the surface potential by

$$V_G' = V_G - V_{FB} = V_{ox} + \psi_s \tag{E.1}$$

where V_{FB} is the flatband voltage. The voltage drop across the oxide V_{ox} is obtained from Poisson's equation

$$\frac{d\mathcal{E}(x,t)}{dx} = \frac{\rho(x)}{\epsilon_{ox}} = -\frac{q}{\epsilon_{ox}} \int_E n_{ot}(E,x,t)dE \tag{E.2}$$

where $n_{ot}(E,x,t)$ is the trapped charge density ($\mathrm{cm}^{-3}\ \mathrm{eV}^{-1}$) and $\mathcal{E}$ is the electric field in the oxide. Integrating from 0 to x,

$$\begin{aligned} \mathcal{E}(x) &= \mathcal{E}(0) - \frac{q}{\epsilon_{ox}} \int_0^x \left[\int_E n_{ot}(E,x,t)dE \right] dx \\ V(x) - V(0) &= -\int_0^x \mathcal{E}(x')dx' \\ &= -\mathcal{E}(0)x + \frac{q}{\epsilon_{ox}} \int_0^x \int_0^{x'} \left[\int_E n_{ot}(E,x'',t)dE \right] dx''dx' \end{aligned} \tag{E.3}$$

From Gauss's law $\epsilon_{ox}\mathcal{E}(0) = \epsilon_s\mathcal{E}_s$ where $\mathcal{E}_s$ is the electric field in the silicon at the interface. Changing the order of integration and defining the positive direction for $\mathcal{E}_s$ is chosen into the silicon, Eq. E.3 can be written

$$\begin{aligned} V_{ox}(t) &= V(t_{ox}) - V(0) \\ &= \frac{\epsilon_s}{\epsilon_{ox}}\mathcal{E}_s t_{ox} + \frac{q}{\epsilon_{ox}} \int_0^{t_{ox}} (t_{ox} - x) \int_E n_{ot}(E,x,t)dEdx \end{aligned} \tag{E.4}$$

The change in gate voltage required to produce the same ψ_s (and $\mathcal{E}_s$) in the presence of the trapped charge is, from Eqs. E.1 and E.4,

$$\Delta V_G = (V_G - V_{G0})|_{\psi_s}$$

$$\begin{aligned} &= \frac{q}{\epsilon_{ox}} \int_0^{t_{ox}} (t_{ox} - x) \int_E n_{ot}(E, x, t) dE dx \\ &= \Delta V_T \end{aligned} \tag{E.5}$$

where V_{G0} is the gate voltage when there is no trapped charge. Equation E.5 is therefore the threshold voltage shift due to the trapped charge, which depends on time through the dynamics of n_{ot}.

Since the observed threshold voltage shifts are, at most, on the order of one millivolt, we shall neglect the second term in Eq. E.4 in the calculation of the Fermi level, the oxide voltage and the band bending in the silicon. For a 1V drop across the oxide this will result in an error of, at most, 0.1%. This term [Eq. E.5] will however, have a very significant effect on the threshold voltage shift in circuits.

In terms of the normalized surface potential

$$\begin{aligned} U(x) &\equiv \frac{E_i(bulk) - E_i(x)}{kT} \\ J_F &\equiv \frac{E_i(bulk) - E_F}{kT} \\ \psi(x) &\equiv \frac{kT}{q} U(x) \end{aligned} \tag{E.6}$$

Equation E.1 becomes

$$V_G - V_{FB} = \frac{kT}{q} U_s + \frac{t_{ox}}{\epsilon_{ox}} \epsilon_s \mathcal{E}_s \tag{E.7}$$

where $U_s = U(0)$ is the normalized surface potential.

The electric field in the silicon at the surface $\mathcal{E}_s$ is obtained from Poisson's equation,

$$\frac{d^2U}{dx^2} = \frac{dU}{dx}\frac{d}{dx}\left(\frac{dU}{dx}\right) = \frac{q^2}{\epsilon_s kT}[n - p + N_A^- - N_D^+]$$

where $\mathcal{E} = U = 0$ in the bulk. Integrating from the surface to x yields the standard result

$$\mathcal{E}_s = \left(\frac{2kT}{\epsilon_s}\right)^{1/2} \left[\int_0^{U_s} \left(n - p + N_A^- - N_D^+\right) dU\right]^{1/2} \mathrm{sgn}(U_s) \tag{E.8}$$

In the Boltzmann limit where $n(U) = n_i \exp(U - U_F)$ and $p(U) = n_i \exp(U_F - U)$, and assuming complete ionization, this can be written in the familiar form

$$\mathcal{E}_s = \left(\frac{2kTn_i}{\epsilon_s}\right)^{1/2} F(U, U_F)\mathrm{sgn}(U_s) \tag{E.9}$$

where

$$F(U, U_F) = \left\{e^{U_F}[e^{-U} + U - 1] + e^{-U_F}[e^U - U - 1]\right\}^{1/2} \tag{E.10}$$

Since the applied pulses will drive the Fermi level close to or into the conduction and valence bands at the surface, Boltzmann statistics cannot be used at the surface. The carrier concentrations are instead calculated from Fermi-Dirac statistics,

$$n = N_c \frac{2}{\sqrt{\pi}} \mathcal{F}_{1/2}(\eta_{fn}) \tag{E.11}$$

$$p = N_v \frac{2}{\sqrt{\pi}} \mathcal{F}_{1/2}(\eta_{fp}) \tag{E.12}$$

where

$$\eta_{fn}(x) \equiv \frac{E_F - E_c(x)}{kT} \tag{E.13}$$

$$\eta_{fp}(x) \equiv \frac{E_v(x) - E_F}{kT} \tag{E.14}$$

N_c and N_v are the effective densities of states in the silicon conduction and valence bands,

$$N_c = 2\left[\frac{2\pi m_n^* kT}{\hbar^2}\right]^{3/2} \tag{E.15}$$

$$N_v = 2\left[\frac{2\pi m_p^* kT}{\hbar^2}\right]^{3/2} \tag{E.16}$$

$\mathcal{F}_{1/2}$ is the Fermi integral of order 1/2,

$$\mathcal{F}_{1/2}(\eta_f) = \int_0^\infty \frac{\eta^{1/2}}{1 + e^{\eta - \eta_f}} d\eta \tag{E.17}$$

Since Eq. E.17 cannot be expressed in closed form, an approximate expression due to Blakemore [246] is used,

$$\mathcal{F}_{1/2}(\eta) \approx \frac{\sqrt{\pi}}{2[c(\eta) + \exp(-\eta)]}$$

where

$$c(\eta) = \frac{3\sqrt{\pi}}{4\{\eta^4 + 50 + 33.6[1 - 0.68\exp(-0.17(x+1)^2)]\}^{3/8}}$$

The fractional error $\Delta\mathcal{F}/\mathcal{F}$ of this approximation is less than $\pm 0.4\%$ over the range $-10 \leq \eta \leq +25$, which is more than adequate for our purposes.

The ionized impurity concentrations are given by [151],

$$N_D^+ = \frac{N_D}{1 + g_D \exp\left(\frac{E_F - E_D}{kT}\right)} = \frac{N_D}{1 + g_D \exp\left(\eta_{fn} + \frac{E_D}{kT}\right)} \tag{E.18}$$

$$N_A^- = \frac{N_A}{1 + g_A \exp\left(\frac{E_A - E_F}{kT}\right)} = \frac{N_A}{1 + g_A \exp\left(\eta_{fp} + \frac{E_A}{kT}\right)} \tag{E.19}$$

where E_A and $E_D \approx 25$meV are the acceptor and donor energies and $g_A \approx 4$ and $g_D \approx 2$ are the corresponding degeneracy factors.

The parameters η_{fn} and η_{fp} are related to U by

$$\eta_{fn} = \frac{E_F - E_c(x)}{kT} = \frac{E_i(bulk) - E_c(x)}{kT} - U_F \tag{E.20}$$

The bulk material is nondegenerate, so

$$E_i(bulk) = \frac{E_c(bulk) + E_v(bulk)}{2} + \frac{3}{4}kT \ln\left(\frac{m_p^*}{m_n^*}\right) \tag{E.21}$$

Substituting Eq. E.21 in Eq. E.20,

$$\eta_{fn} = U(x) - U_F - \frac{E_G}{2kT} + \frac{3}{4}\ln\left(\frac{m_p^*}{m_n^*}\right) \tag{E.22}$$

Since $\eta_{fn} + \eta_{fp} = -E_G/kT$,

$$\eta_{fp} = -U(x) + U_F - \frac{E_G}{2kT} - \frac{3}{4}\ln\left(\frac{m_p^*}{m_n^*}\right) \tag{E.23}$$

The bulk Fermi level U_F is obtained from the condition of charge neutrality in the bulk

$$p - n + N_D^+ - N_A^- = 0 \tag{E.24}$$

Since the bulk is nondegenerate, Boltzmann approximations, may be used for n and p in Eq. E.24. Alternatively, Eq. E.24 may be solved iteratively for U_F using Eqs. E.11, E.12, E.18 and E.19.

Having obtained U_F, Eqs. E.7 and E.8 are then solved iteratively for the normalized surface potential U_s at the given gate voltage, using Eqs. E.11, E.12, E.18, E.19, E.22 and E.23.

The Fermi level at the interface is then given by Eq. E.22 at $x = 0$,

$$\frac{E_F - E_c(0)}{kT} = U_s - U_F - \frac{E_G}{2kT} + \frac{3}{4}\ln\left(\frac{m_p^*}{m_n^*}\right) \tag{E.25}$$

Appendix F

Recovery Voltage for Donor-like Traps

In this appendix we we repeat the analysis of Sec. 5.4.1 for donor-like traps, i.e. neutral when occupied by an electron and positively charged when empty. It is shown that the previous results, Eqs. 5.11 and 5.16, hold regardless of acceptor or donor-like nature of the traps.

Trapping from the Bottom Plate

$\underline{V_p > 0}$

Donor-type traps become filled during positive stress and are therefore neutral at the beginning of the float period. The traps emit electrons during the float period, becoming positively charged. The analysis of Fig. 5.4 applies to this case with the substitution $\Delta n_{ot} \rightarrow \Delta p_{ot}$ and the polarities of the charge density, electric field and potential reversed. Thus,

$$\begin{aligned} \delta V_r^b(t_r) &= \frac{qx}{\epsilon_{ox}} \Delta p_{ot}(E_t, x, t_r) \Delta E_t \Delta x \\ &= -\frac{qx}{\epsilon_{ox}} \Delta n_{ot}(E_t, x, t_r) \Delta E_t \Delta x \end{aligned} \tag{F.1}$$

which is identical to Eq. 5.12. The derivation again leads to Eq. 5.11. The recovery voltage is positive since $f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x) > 0$.

$\underline{V_p < 0}$

In this case, traps empty during stress and are positively charged at the beginning of the float period. As traps fill they become neutral and the analysis of Fig. 5.3 applies

with $\Delta n_{ot} \rightarrow \Delta p_{ot}$ and all polarities reversed.

$$\begin{aligned} \delta V_r^b &= -\frac{qx}{\epsilon_{ox}}[\Delta p_{ot}(E_t, x, 0) - \Delta p_{ot}(E_t, x, t_r)]\Delta E_t \Delta x \\ &= \frac{qx}{\epsilon_{ox}}[\Delta n_{ot}(E_t, x, 0) - \Delta n_{ot}(E_t, x, t_r)]\Delta E_t \Delta x \end{aligned} \tag{F.2}$$

which is identical to Eq. 5.8 and Eq. 5.11 again results. Since $f(E_t + q\mathcal{E}x) - f_0(E_t + q\mathcal{E}_o x) < 0$ for $V_p < 0$, the recovery voltage is negative.

Trapping from the Top Plate

$\underline{V_p > 0}$

Donor traps near the top plate empty during positive stress, becoming positively charged. During the float period they refill, becoming neutral. The analysis of Fig. 5.6 applies, with $\Delta n_{ot} \rightarrow \Delta p_{ot}$ and all polarities reversed. The contribution to the recovery voltage is therefore

$$\begin{aligned} \delta V_r^t &= \frac{qx}{\epsilon_{ox}}[\Delta p_{ot}(E_t, x, 0) - \Delta p_{ot}(E_t, x, t_r)] \\ &= -qx\epsilon_{ox}[\Delta n_{ot}(E_t, x, 0) - \Delta n_{ot}(E_t, x, t_r)] \end{aligned} \tag{F.3}$$

This is identical to Eq. 5.17 and again leads to Eq. 5.16. Since $f(E_t - q\mathcal{E}x) - f_0(E_t - q\mathcal{E}_o x) < 0$, the recovery voltage is positive.

$\underline{V_p < 0}$

Under negative stress, traps fill and become neutral. During the float period, they empty becoming positively charged. The analysis of Fig. 5.5 applies, with $\Delta n_{ot} \rightarrow \Delta p_{ot}$ and all polarities reversed. Thus,

$$\begin{aligned} \delta V_r^t &= -\frac{qx}{\epsilon_{ox}}\Delta p_{ot}(E_t, x, t_r)\Delta E_t \Delta x \\ &= \frac{qx}{\epsilon_{ox}}\Delta n_{ot}(E_t, x, t_r)\Delta E_t \Delta x \end{aligned} \tag{F.4}$$

which is identical to Eq. 5.15 so that the recovery voltage is given by Eq. 5.16. Since $f(E_t - q\mathcal{E}x) - f_0(E_t - q\mathcal{E}_o x) > 0$ for $V_p < 0$, the recovery voltage is negative. Thus,

regardless of the polarity of the charging voltage or whether the traps are acceptor-type or donor-type, the contribution to the recovery voltage due to trapping from the

bottom plate V_r^b is given by Eq. 5.11 and the contribution from the top plate V_r^t is given by Eq. 5.16. The total recovery voltage is the superposition of the contributions due to each plate

$$V_r = V_r^b + V_r^t \tag{F.5}$$

As shown above, the polarities the components are such that the contributions from each plate add constructively.

Appendix G

List of Symbols

A	area (cm^2)
c_{ns}	interface-state electron-capture coefficient (cm^3/s)
c_{ps}	interface-state hole-capture coefficient (cm^3/s)
C_{ox}	oxide capacitance per unit area (F/m^2)
D_{it}	interface state density (cm^{-2} eV^{-1})
D_{ot}	oxide trap density (cm^{-3} eV^{-1})
D_{ot}^{L}	oxide trap density bordering on bird's beak (cm^{-1})
D_{ot}^{W}	oxide trap density bordering on source-drain regions (cm^{-1})
$\mathcal{E}$	electric field (V/cm)
$\mathcal{E}_o$	electric field at initial bias condition (V/cm)
$\mathcal{E}_s$	electric field in the semiconductor at the surface (V/cm)
E	electron energy measured upward from silicon conduction band (eV)
E_A	acceptor ionization energy
E_c	lowest conduction band energy in the semiconductor (eV)
E_{co}	lowest conduction band energy in the insulator (eV)
E_D	donor ionization energy
E_F	Fermi energy or Fermi level (eV)
E_{F0}	Fermi energy or Fermi level at initial bias condition (eV)
E_G	Energy gap in the semiconductor (eV)
E_{Gb}	Energy gap in the insulator (eV)
E_t	Trap energy with respect to oxide bands (eV)
E_v	highest valence band energy in the semiconductor (eV)
E_{vo}	highest valence band energy in the insulator (eV)
$\mathcal{F}_{1/2}$	Fermi integral of order 1/2; Eq. E.17
f	Fermi function, defined in Eq. 3.75.
f_0	Fermi function at initial bias condition
f_{it}	Non-steady-state occupancy function for a trap at the interfacpe, $x = 0$.
f_l	Fermi factor for a state on the left side of a barrier region

f_{ot}	Non-steady-state occupancy function for a trap at depth x in the oxide.
f_r	Fermi factor for a state on the right side of a barrier region
F	$= q\mathcal{E}$ (eV)
F_{it}	interface state quasi-Fermi level (eV)
F_n	electron quasi-Fermi level (eV)
F_p	hole quasi-Fermi level (eV)
g_A	acceptor degeneracy factor
g_D	donor degeneracy factor
$\hbar$	Planck's constant
H_l	Hamiltonian for left-side region in Bardeen's method
H_r	Hamiltonian for right-side region in Bardeen's method
I_{DS}	drain current (A)
k	Boltzmann's constant (8.617×10^{-5} eV/K)
k_x	wavevector component in the semiconductor perpendicular to the interface, defined in Eq. B.14 (cm^{-1}).
k_y	wavevector component in the semiconductor parallel to the interface (cm^{-1})
k_z	wavevector component in the semiconductor parallel to the interface (cm^{-1})
$\vec{k}_{\|\|}$	wavevector component in the semiconductor parallel to the interface (cm^{-1})
K	imaginary wavevector in the insulator; defined in Eq. B.25 (cm^{-1}).
K_c	imaginary wavevector in the insulator in a single-band model with tunneling controlled by the conduction band (cm^{-1})
K_x	imaginary wavevector in the insulator, perpendicular to the interface; defined in Eq. B.13 (cm^{-1}).
K_v	imaginary wavevector in the insulator in a single-band model with tunneling controlled by the valence band (cm^{-1})
L	channel length (cm)
L_x	length of semiconductor region perpendicular to the interface (cm)
L_y	lateral dimension of semiconductor region parallel to the interface (cm)
L_z	lateral dimension of semiconductor region parallel to the interface (cm)
m_c^*	conduction band effective mass in the semiconductor (g)
m_l	effective mass on the left side of a potential barrier
m_o	free electron rest mass (g)
m_r	effective mass on the right side of a potential barrier
m_v^*	valence band effective mass in the semiconductor (g)
M_{lr}	matrix element for tunneling from a state on the left to a state on the right side of a barrier
M^*	effective mass in the oxide (g)
M_c^*	conduction band effective mass in the insulator (g)
M_v^*	valence band effective mass in the insulator (g)

n_1	defined in Eq. 3.53.
n_i	intrinsic carrier concentration (cm^{-3})
n_{it}	density of filled interface states (cm^{-2} eV^{-1})
n_{ot}	density of filled oxide traps (cm^{-3} eV^{-1})
n_s	electron concentration at the surface (cm^{-3})
n_{s0}	electron concentration at the surface in equilibrium (cm^{-3})
N_A	acceptor concentration (cm^{-3})
N_c	effective conduction band density of states; Eq. E.15 (cm^{-3})
N_D	donor concentration
N_{ot}	$= \int D_{ot} dE$ (cm^{-3})
N_v	effective valence band density of states; Eq. E.16 (cm^{-3})
p_1	defined in Eq. 3.54.
p_{it}	$= D_{it} - n_{it}$, density of unoccupied interface states ($cm^{-2}eV^{-1}$)
p_{ot}	$= D_{ot} - n_{ot}$, density of unoccupied oxide traps ($cm^{-3}eV^{-1}$)
p_s	hole concentration at the surface (cm^{-3})
p_{s0}	hole concentration at the surface in equilibrium (cm^{-3})
q	magnitude of electron charge (1.602×10^{-19} coul)
Q	lattice normal coordinate
r_{lr}	rate at which electrons of a given energy and parallel wavevector tunnel from states on the left to states on the right of a barrier region (s^{-1})
r_{rl}	rate at which electrons of a given energy and parallel wavevector tunnel from states on the left to states on the right of a barrier region (s^{-1})
Q_b	depletion charge ($coul/cm^2$)
Q_N	inversion layer charge [Eq. 4.2] ($coul/cm^2$)
Q_{ot}	oxide trapped charge ($coul/cm^2$)
t_d	discharge (short-circuit) time in dielectric relaxation measurement (s)
t_{ox}	oxide thickness (cm)
t_r	relaxation time (s)
t_s	stress time (s)
T	temperature (K)
$T_{0\to x}$	tunneling rate from the interface to an oxide trap at depth x (cm^2eVs^{-1})
T_D	Debye temperature (K)
$T(x)$	tunneling rate (cm^2eVs^{-1})
$T_{x\to 0}$	tunneling rate from an oxide trap at depth x to the interface (cm^2eVs^{-1})
U	normalized potential
U_s	normalized surface potential
$\overline{v}$	thermal velocity (cm)
U_c	insulator conduction band potential energy (eV)
V_δ	Delta function potential strength
V_{DS}	$= V_D - V_S$ drain-source voltage (V)
V_{FB}	= flatband voltage (V)
V_G	= gate voltage (V)

$V_G{}'$	$= V_G - V_{FB}$ (V)
V_{GS}	$= V_G - V_S$ gate-source voltage (V)
V_{GS}^o	= initial gate-source voltage (V)
$V_{GS}\|_{acc}$	= Gate-source voltage at which accumulation occurs (V)
V_{ox}	voltage drop across the oxide (V)
V_p	amplitude of voltage pulse during stress (V)
V_T	threshold voltage (V)
U_v	insulator valence band potential energy (eV)
W	channel length (cm)
W_{lr}	Tunneling transistion probability per unit time from a single occupied state on the left to a set of unoccupied states at the same energy on the right side of a potential barrier ($s^{-1}cm^{-3}$)
x	depth into oxide, measured from the interface (cm)
x_o	characteristic tunneling depth; defined in Eq. 2.3
α	defined by Eq. 4.86 ($s^{-1}V^{-1}$)
β	$= \mu C_{ox} W/L$ (A/V^2)
γ	body-effect parameter ($V^{1/2}$) [Eq. 7.3]
γ_l	defined in Eq. 4.95
γ_w	defined in Eq. 4.94
ΔE_B	Barrier height for capture in lattice relaxation multiphonon emission (eV)
ΔE_{Bc}	Barrier height for capture of an electron from the silicon conduction band by an oxide trap, in the lattice relaxation model (eV)
ΔE_{Bv}	Barrier height for capture of a hole from the silicon valence band by an oxide trap, in the lattice relaxation model (eV)
ΔV	$= V_{GS} - V_T$ (V)
ΔV_{GS}	equivalent gate-source voltage shift (V)
ΔV_T	threshold voltage shift (V)
$\Delta V_T'$	effective threshold voltage shift (including correlated mobility fluctuations) (V)
ϵ_{ox}	permittivity of the oxide (F/cm)
ϵ_{si}	permittivity of silicon (F/cm)
η_{fn}	defined by Eq. E.13
η_{fp}	defined by Eq. E.14
κ	imaginary wavevector in the semiconductor [Eq. 3.95] (cm^{-1})
κ'	defined by Eq. (cm^{-1})
λ	the WKB integral (cm^{-1}), defined by Eq. 3.90.
μ_n	electron mobility (cm^2/Vs)
ρ_l	density of states on left side of tunneling barrier
ρ_r	density of states on right side of tunneling barrier
σ_{no}	capture cross section of an interface state for an electron (cm^2)
σ_n	capture cross section of an oxide trap for an electron (cm^2)

σ_{po}	capture cross section of an interface state for a hole (cm^2)
σ_p	capture cross section of an oxide trap for a hole (cm^2)
τ_o	tunneling time constant prefactor (seconds)
τ_{it}	thermal time constant for interface states [Eq. 3.100]
τ_{it}^o	thermal time constant for interface states at intitial bias condition
τ_{ot}	$= 1/\omega$; time constant for tunneling transitions to and from an oxide trap
$\tau_{ot/cb}$	time constant for tunneling between oxide traps and conduction band states during stress.
$\tau_{ot/cb}^o$	time constant for tunneling between oxide traps and conduction band states, before and after stress.
$\tau_{ot/vb}$	time constant for tunneling between oxide traps and valence band states during stress.
$\tau_{ot/vb}^o$	time constant for tunneling between oxide traps and valence band states, before and after stress.
$\tau_{ot/it}$	time constant for tunneling between oxide traps and interface states during stress.
$\tau_{ot/it}^o$	time constant for tunneling between oxide traps and interface states, before and after stress.
ϕ_{acc}	surface potential at the interface at the onset of accumulation (V)
ϕ_F	bulk Fermi level
ϕ_c	barrier height between oxide and silicon conduction bands at the interface (eV)
ϕ_v	barrier height between oxide and silicon valence bands at the interface (eV)
ψ_l	wavefunction of electron in left-side region in Bardeen's method
ψ_r	wavefunction of electron in right-side region in Bardeen's method
ω	$= 1/\tau_{ot}$; tunneling rate to and from an oxide trap
$\omega_{ot/cb}$	$= 1/\tau_{ot/cb}$; rate of tunneling from a conduction band state into an oxide trap (s^{-1})
$\omega_{ot/it}$	$= 1/\tau_{ot/it}$; rate of tunneling from an interface state into an oxide trap (s^{-1})
$\omega_{ot/vb}$	$= 1/\tau_{ot/vb}$; rate of tunneling from a valence band state into an oxide trap (s^{-1})
Ω	$= L_x L_y L_z$; volume of semiconductor region (cm^3)

Appendix H

Abbreviations and Acronyms

A/D	analog-to-digital
AEI	avalanche electron injection
cb	conduction band
CC-DLTS	constant capacitance deep level transient spectroscopy
CV	capacitance-voltage
D/A	digital-to-analog
DIGS	disorder-induced gap state
DLTS	deep level transient spectroscopy
DUT	device under test
FNTEI	Fowler-Nordheim tunneling electron injection
it	interface state (trap at $x = 0$)
IPE	internal photoemission
keV-EBG	kilo-volt electron beam generation
lr	lattice relaxation
LRME	lattice relaxation multiphonon emission
MIM	metal-insulator-metal
MISFET	metal-insulator-semiconductor field-effect transistor
MNOS	metal-nitride-oxide-semiconductor
MOSFET	metal-oxide-semiconductor field-effect transistor
MPE	multiphonon emission
ot	oxide trap
SRH	Shockley-Read-Hall (analysis)
vb	valence band
VUVEI	vacuum ultra-violet light electron injection

Bibliography

[1] C. T. Sah, "VLSI device reliability modeling," in *Proc. Int. VLSI-TSA Conf.*, pp. 153–162, 1987.

[2] A. L. McWhorter, *1/f Noise and Related Surface Effects in Germanium.* PhD thesis, Massachusetts Institute of Technology, Cambridge, 1955.

[3] I. A. Young, P. R. Gray, and D. A. Hodges, "Analog NMOS sampled data recursive filters," in *ISSCC Dig. Tech. Papers*, pp. 156–157, Feb. 1977.

[4] R. Castello and P. R. Gray, "A high-performance micropower switched-capacitor filter," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 1122–1132, Dec. 1985.

[5] R. W. Brodersen, P. R. Gray, and D. A. Hodges, "MOS switched-capacitor filters," *Proc. IEEE*, vol. 67, pp. 61–75, Jan. 1979.

[6] D. J. Allstot, R. W. Brodersen, and P. R. Gray, "MOS switched capacitor ladder filters," *IEEE J. Solid-State Circuits*, vol. SC-13, no. 6, pp. 806–814, Dec. 1978.

[7] H. S. Lee, D. A. Hodges, and P. R. Gray, "A self-calibrating 15 bit CMOS A/D converter," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 813–819, Dec. 1984.

[8] P. W. Li, M. J. Chin, P. R. Gray, and R. Castello, "A ratio independent algorithmic analog-to-digital conversion technique," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 828–836, Dec. 1984.

[9] J. L. McCreary and P. R. Gray, "All-MOS charge redistribution analog-to-digital conversion techniques – part I," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 371–379, December 1975.

[10] T. L. Tewksbury, H. S. Lee, and G. A. Miller, "The effects of oxide traps on the large-signal transient response of analog MOS circuits," *IEEE J. Solid-State Circuits*, vol. 24, no. 2, pp. 542–544, April 1989.

[11] K. S. Tan, S. Kiriaki, M. de Wit, J. W. Fattaruso, C. Y. Tsay, W. E. Matthews, and R. K. Hester, "Error correction techniques for high-performance differential A/D converters," *IEEE J. Solid-State Circuits*, vol. 25, no. 6, pp. 1328–1327, Dec. 1990.

[12] B. J. Sheu, J. Shieh, and M. Patel, "Modeling charge injection in MOS analog switches," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 214–216, Feb. 1987.

[13] J. Shieh, M. Patel, and B. J. Sheu, "Measurement and analysis of charge injection in MOS analog switches," *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 277–281, April 1987.

[14] B. J. Sheu and C. Hu, "Switch-induced error voltage on a switched capacitor," *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 519–525, Aug. 1984.

[15] H. S. Lee and D. A. Hodges, "Self-calibration technique for A/D converters," *IEEE Trans. Circuits Syst.*, vol. CAS-30, pp. 188–190, March 1983.

[16] R. J. Kansy, "Response of correlated double sampling circuit to 1/f noise," *IEEE J. Solid-State Circuits*, vol. SC-15, pp. 373–375, June 1980.

[17] H. S. Lee and D. A. Hodges, "A precision measurement technique for residual polarization in integrated circuit capacitors," *IEEE Electron Device Lett.*, vol. EDL-5, no. 10, pp. 417–420, October 1984.

[18] M. Declercq and P. Jespers, "Analysis of interface properties in MOS transistors by means of charge pumping measurements," *Rev. HF Acta. Techn. Belgica.*, vol. 9, pp. 244–253, 1974.

[19] D. Vuillaume and J. C. Bourgoin, "Transient spectroscopy of Si-SiO_2 interface states," *Surface Science*, vol. 162, pp. 680–686, 1985.

[20] N. F. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystals.* Oxford University Press, London, 1940.

[21] E. H. Snow, A. S. Grove, B. E. Deal, and C. T. Sah, "Ion transport phenomena in insulating films," *J. Appl. Phys.*, vol. 36, no. 5, pp. 1664–1673, May 1965.

[22] S. R. Hofstein, "An investigation of instability and charge motion in metal-silicon oxide-silicon structures," *IEEE Trans. Electron Devices*, vol. 13, no. 12, pp. 222–237, Feb. 1966.

[23] P. Debye, *Polar Molecules.* New York: Chemical Catalog Company, 1929.

[24] J. C. Maxwell, *Electricity and Magnetism*, vol. 1. London, England: Oxford University Press, 1892.

[25] K. W. Wagner, "Zur theorie der unvollkommenen dielektrika," *Ann. der Physik*, vol. 40, no. 5, pp. 817–855, 1913.

[26] E. H. Snow and B. E. Deal, "Polarization phenomena and other properties of phosphosilicate glass films on silicon," *J. Electrochem. Soc.*, vol. 113, no. 3, pp. 263–269, March 1966.

[27] P. Rossel, H. Martinot, and A. Serhal, "Instabilites temporelles des transistors MOS–I," *Solid-St. Electron*, vol. 13, pp. 257–268, 1970.

[28] W. Shockley and W. Read, "Statistics of the recombination of holes and electrons," *Phys. Rev.*, vol. 87, no. 835, pp. 835–842, 1952.

[29] S. Christensson, I. Lundstrom, and C. Svensson, "Low frequency noise in MOS transistors," *Solid-St. Electron*, vol. 11, pp. 797–812, 1968.

[30] M. E. Zvanut, F. J. Feigl, and J. D. Zook, "A defect relaxation model for bias instabilities in metal-oxide-semiconductor capacitors," *J. Appl. Phys.*, vol. 64, no. 4, pp. 2221–23, Aug. 1988.

[31] D. J. Breed, "Non-ionic room temperature instabilities in MOS devices," *Solid-St. Electron*, vol. 17, pp. 1229–1243, 1974.

[32] S. A. Abbas and R. C. Dockerty, "Hot-carrier instability in IGFET's," *Appl. Phys. Lett.*, vol. 27, no. 3, pp. 147–148, Aug. 1975.

[33] R. B. Fair and R. C. Sun, "Threshold-voltage instability in MOSFET's due to channel hot-hole emission," *IEEE Trans. Electron Devices*, vol. ED-28, no. 1, pp. 83–94, Jan. 1981.

[34] C. Hu. S. C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, and K. W. Terrill, "Hot-electron-induced MOSFET degradation– model, monitor, and improvement," *IEEE Trans. Electron Devices*, vol. 32, no. 2, pp. 375–385, Feb. 1985.

[35] T. Tsuchiya, T. Kobayashi, and S. Nakajima, "Hot carrier degradation mechanism in Si nMOSFET's," in *Extended Abstracts of the 17th Conference on Solid State Devices and Materials, Tokyo*, pp. 21–24, 1985.

[36] S. K. Lai and D. R. Young, "Effects of avalanche injection of electrons into silicon-dioxide– generation of fast and slow interface states," *J. Appl. Phys.*, vol. 52, no. 10, pp. 6231–6240, Oct. 1981.

[37] F.-C. Hsu and S. Tam, "Relationship between MOSFET degradation and hot-electron- induced interface-state generation," *IEEE Electron Device Lett.*, vol. EDL-8, no. 2, pp. 50–52, Feb. 1984.

[38] B. S. Doyle, M. Bourcerie, J.-C. Marchetaux, and A. Boudou, "Interface state creation and charge trapping in the medium-to-high gate voltage range ($v_d/2 \geq v_g \geq v_d$) during hot-carrier stressing of n-MOS transistors," *IEEE Trans. Electron Devices*, vol. 37, no. 3, pp. 744–754, March 1990.

[39] S. K. Lai, "Two-carrier nature of interface-state generation in hole trapping and radiation damage," *Appl. Phys. Lett.*, vol. 39, no. 1, pp. 58–60, July 1981.

[40] B. S. Doyle, M. Bourcerie, J.-C. Marchetaux, and A. Boudou, "Dynamic channel hot-carrier degradation of NMOS transistors by enhanced electron-hole injection into the oxide," *IEEE Electron Device Lett.*, vol. EDL-8, no. 5, pp. 237–239, May 1987.

[41] M. Bourcerie, B. S. Doyle, J. C. Marchetaux, J. C. Soret, and A. Boudou, "Relaxable damage in hot-carrier stressing of n-MOS transistors–oxide traps in the near interfacial region of the gate oxide," *IEEE Trans. Electron Devices*, vol. 37, no. 3, pp. 708–717, March 1990.

[42] B. S. Doyle, M. Bourcerie, J.-C. Marchetaux, and A. Boudou, "Relaxation effects in NMOS transistors after hot-carrier stressing," *IEEE Electron Device Lett.*, vol. EDL-8, no. 5, pp. 234–236, May 1987.

[43] K. Mistry and B. S. Doyle, "The role of electron trap creation in enhanced hot-carrier degradation during ac stress," *IEEE Electron Device Lett.*, vol. 11, no. 6, pp. 267–269, June 1990.

[44] H. Gesch, J.-P. Leburton, and G. E. Dorda, "Generation of interface states by hot hole injection in MOSFET's," *IEEE Trans. Electron Devices*, vol. ED-29, no. 5, pp. 913–918, May 1982.

[45] P. Cuevas, "A simple explanation for the apparent relaxation effect associated with hot-carrier phenomenon in MOSFETs," *IEEE Electron Device Lett.*, vol. 9, no. 12, pp. 627–629, Dec. 1988.

[46] H. S. Fu and C. T. Sah, "Theory and experiments on surface 1/f noise," *IEEE Trans. Electron Devices*, vol. ED-19, no. 2, pp. 273–285, Feb. 1972.

[47] P. O. Lauritzen, "Low-frequency generation noise in junction field effect transistors," *Solid-St. Electron*, vol. 8, p. 41, 1965.

[48] F. M. Klaassen, "Characterization of low 1/f noise in MOS transistors," *IEEE Trans. Electron Devices*, vol. ED-18, pp. 887–891, Oct. 1971.

[49] S. T. Hsu, "Surface state related 1/f noise in MOS transistors," *Solid-St. Electron*, vol. 13, pp. 1451–1459, 1970.

[50] F. Berz, "Theory of low-frequency noise in Si MOSTs," *Solid-St. Electron*, vol. 13, pp. 631–647, 1970.

[51] R. Jayaraman and C. G. Sodini, "A 1/f noise technique to extract the oxide trap density near the conduction band edge of silicon," tech. rep., M.I.T., Dec. 1988. VLSI Memo No. 88-499.

[52] Z. Celik-Butler and T. Y. Hsiang, "Determination of Si-SiO_2 interface trap density by 1/f noise measurements," *IEEE Trans. Electron Devices*, vol. 35, no. 10, pp. 1651–1655, Oct. 1988.

[53] D. Vuillaume, "Comments on determination of Si-SiO_2 interface trap density by 1/f noise measurements," *IEEE Trans. Electron Devices*, vol. 37, pp. 824–825, March 1990.

[54] Z. Celik-Butler and T. Y. Hsiang, "Reply to comments on determination of Si-SiO_2 interface trap density by 1/f noise measurements," *IEEE Trans. Electron Devices*, vol. 37, pp. 825–826, March 1990.

[55] H. Preier, "Contributions of surface states to MOS impedance," *Appl. Phys. Lett.*, vol. 10, no. 12, pp. 361–363, June 1967.

[56] B. Balland and P. Pinard, "Interfacial tunnel interaction in MOS elements in depletion," *Phys. Stat. Sol. (A)*, vol. 47, pp. 251–258, 1978.

[57] D. Sands, K. M. Brunson, and C. B. Thomas, "The relative contribution of recombination and tunnelling at interface states to the a.c. conductance of metal-insulator-semiconductor diodes," *Solid-St. Electron*, vol. 30, no. 5, pp. 543–548, 1987.

[58] F. P. Heiman and G. Warfield, "The effects of oxide traps on the MOS capacitance," *IEEE Trans. Electron Devices*, vol. ED-12, pp. 167–178, April 1965.

[59] I. Lundstrom, S. Christensson, and C. Svensson, "Hysteresis effects in MOS transistors due to carrier trapping at the semiconductor-oxide interface," in *Physical Aspects of Noise in Electronic Circuits*, pp. 131–134, Institute of Physics and The Physical Society, Sept. 1968.

[60] I. Lundstrom, S. Christensson, and C. Svensson, "Carrier trapping hysteresis in MOS transistors," *Phys. Stat. Sol. (A)*, vol. 1, no. 395, pp. 395–407, 1970.

[61] P. Rossel, H. Martinot, and A. Serhal, "Instabilites temporelles des transistors MOS–II," *Solid-St. Electron*, vol. 13, pp. 425–438, 1970.

[62] T. L. Tewksbury, "Characterization of nonidealities in integrated circuit transistors and capacitors," Master's thesis, Massachusetts Institute of Technology, Cambridge, 1987.

[63] J. S. Brugler and P. G. A. Jespers, "Charge pumping in MOS devices," *IEEE Trans. Electron Devices*, vol. ED-16, no. 3, pp. 297–302, March 1969.

[64] G. Groeseneken, H. E. Maes, N. Beltran, and R. R. D. Keersmaecker, "A reliable approach to charge-pumping measurements in MOS transistors," *IEEE Trans. Electron Devices*, vol. ED-31, no. 1, pp. 42–53, Jan. 1984.

[65] N. S. Saks and M. G. Ancona, "Determination of interface trap capture cross sections using three-level charge pumping," *IEEE Electron Device Lett.*, vol. 11, no. 8, pp. 339–341, Aug. 1990.

[66] W. L. Tseng, "A new charge pumping method of measuring Si-SiO_2 interface states," *J. Appl. Phys.*, vol. 62, no. 2, pp. 591–599, July 1987.

[67] U. Cilingiroglu, "A general model for interface-trap charge-pumping effects in MOS devices," *Solid-St. Electron*, vol. 28, no. 11, pp. 1127–1141, 1985.

[68] A. B. M. Elliot, "The use of charge pumping currents to measure surface state densities in MOS transistors," *Solid-St. Electron*, vol. 19, pp. 241–247, 1976.

[69] U. Cilingiroglu, "A pulsed interface-probing technique for MOS interface characterization at mid-gap levels," *IEEE Trans. Electron Devices*, vol. 35, no. 12, pp. 2391–96, Dec. 1988.

[70] X. M. Li and M. J. Deen, "A new charge pumping method for determining the spatial interface state density distribution in MOSFETs," in *IEDM*, pp. 85–87, 1990.

[71] P. Heremans, J. Witters, G. Groeseneken, and H. E. Maes, "Analysis of the charge pumping technique and its application for the evaluation of MOSFET degradation," *IEEE Trans. Electron Devices*, vol. 36, no. 7, pp. 1318–35, July 1989.

[72] D. Vuillaume, J. C. Bourgoin, and M. Lannoo, "Oxide traps in Si-SiO_2 structures characterized by tunnel emission with deep-level transient spectroscopy," *Phys. Rev. B*, vol. 34, no. 2, pp. 1171–83, July 1986.

[73] H. Lakhardi, D. Vuillaume, and J. C. Bourgoin, "Spatial and energetic distribution of Si-SiO_2 near-interface states," *Phys. Rev. B*, vol. 38, no. 18, pp. 13124–32, Dec. 1988.

[74] J. A. V. Vechten, "Consequences of anion vacancy nearest-neighbor hopping in III-V compound semiconductors: Drift in InP metal-insulator-semiconductor field effect transistors," *J. Appl. Phys.*, vol. 57, no. 6, pp. 1956–60, March 1985.

[75] M. T. Juang, J. F. Wager, and J. A. V. Vechten, "Phosphorous vacancy nearest-neighbor hopping induced instabilities in InP capacitors," in *Dielectric Films on Compound Semiconductors*, pp. 180–186, The Electrochemical Society, Inc., 1988. Volume 88-15.

[76] H. Hasegawa, H. Masuda, L. He, J. K. Luo, T. Sawada, and H. Ohno, "Transient behavior of interface state continuum at InP insulator-semiconductor interface," in *Dielectric Films on Compound Semiconductors*, pp. 203–216, The Electrochemical Society, Inc., 1988. Vol. 88-15.

[77] H. Hasegawa and H. Ohno, "Unified disorder induced gap state model for insulator-semiconductor and metal semiconductor interfaces," *J. Vac. Sci. Technol. B*, vol. 4, no. 4, pp. 1130–1138, Jul/Aug 1986.

[78] H. Hasegawa, L. He, H. Ohno, T. Sawada, T. Haga, Y. Abe, and H. Takahashi, "Electronic and microstructural properties of disorder-induced gap states at compound semiconductor-insulator interfaces," *J. Vac. Sci. Technol. B*, vol. 5, no. 4, pp. 1097–1107, Jul/Aug 1987.

[79] H. Koelmans and H. C. D. Graaff, "Drift phenomena in CdSe thin film FET's," *Solid-St. Electron*, vol. 10, pp. 997–1005, 1967.

[80] M. Okamura and T. Kobayashi, "Slow current-drift mechanism in n-channel inversion type InP-MISFET," *Jpn. J. Appl. Phys.*, vol. 19, no. 11, pp. 2143–2150, Nov. 1980.

[81] E. Yamaguchi, M. Minakata, and Y. Furukawa, "Current drift phenomena and spectroscopic measurement method for insulator trap level parameters in InP MISFETs," *Jpn. J. Appl. Phys.*, vol. 23, no. 1, pp. L49–L51, Jan. 1984.

[82] D. L. Lile and M. J. Taylor, "The effect of interfacial traps on the stability of insulated gate device on InP," *J. Appl. Phys.*, vol. 54, no. 1, pp. 260–267, Jan. 1983.

[83] S. M. Goodnick, T. Hwang, and C. W. Wilmsen, "New model for slow current drift in InP metal-insulator- semiconductor field-effect transistors," *Appl. Phys. Lett.*, vol. 44, no. 4, pp. 453–455, Feb. 1984.

[84] P. van Staa, H. Rombach, and R. Kassing, "Time-dependent response of interface states in indium phosphide metal-insulator-semiconductor capacitors investigated with constant-capacitance deep-level transient spectroscopy," *J. Appl. Phys.*, vol. 54, no. 7, pp. 4014–21, July 1983.

[85] S. J. Prasad and S. J. T. Owen, "The InP-SiO_2 interface: Electron tunneling into oxide traps," in *Dielectric Films on Compound Semiconductors*, pp. 23–32, The Electrochemical Society, Inc., 1985. Vol. 86-3.

[86] S. J. Prasad, *A Study of the Effects of Oxide Traps on the Characteristics of Indium-Phosphide MISFETs.* PhD thesis, Oregon State University, 1985.

[87] H. Hasegawa, T. Sawada, and H. Ohno, "Mechanism and suppression of drain current drift in InP MISFETs," in *Int. Symp. GaAs and Related Compounds, Biarritz*, pp. 569–574, 1984. Inst. Phys. Conf. Ser. No. 74.

[88] H. Sewell and J. C. Anderson, "Slow states in InSb/SiO_x thin film transistors," *Solid-St. Electron*, vol. 18, pp. 641–649, 1975.

[89] M. Pulver and G. Dorda, "Flat-band voltage hysteresis in MNOS structures," *Phys. Stat. Sol. (A)*, vol. 1, no. 65, pp. 65–70, 1970.

[90] G. Dorda and M. Pulver, "Tunnel mechanism in MNOS structures," *Phys. Stat. Sol. (A)*, vol. 1, no. 71, pp. 71–79, 1970.

[91] C. T. Sah, "Models and experiments on degradation of oxidized silicon," *Solid-St. Electron*, vol. 33, no. 2, pp. 147–167, 1990.

[92] D. J. DiMaria, "Defects and impurities in thermal SiO_2," in *The Physics of SiO_2 and its Interfaces* (S. T. Pantelides, ed.), pp. 160–178, Pergamon Press, 1978.

[93] M. Xu, C. Tan, and Y. Wang, "Oxide current relaxation spectroscopy in tunneling metal-oxide-semiconductor structures under high field stress," *J. Appl. Phys.*, vol. 67, no. 11, pp. 6924–6929, 1990.

[94] C. T. Sah, "The equivalent circuit model in solid-state electronics – part II: The multiple energy level defect centers," *Proc. of the IEEE*, vol. 55, no. 5, pp. 672–684, May 1967.

[95] M. Lax, "Cascade capture of electrons in solids," *Phys. Rev.*, vol. 119, no. 5, pp. 1502–1523, Sept. 1960.

[96] F. L. Terry, P. W. Wyatt, M. L. Naiman, B. P. Mathur, and C. T. Kirk, "High-field electron capture and emission in nitrided oxides," *J. Appl. Phys.*, vol. 57, pp. 2036–2039, March 1985.

[97] R. H. Fowler and L. Nordheim, "Electron emission in intense electric fields," *Proc. Roy. Soc., Lond. Ser.*, vol. A119, pp. 173–181, 1928.

[98] Z. A. Weinberg, "On tunneling in metal-oxide-silicon structures," *J. Appl. Phys.*, vol. 53, no. 7, pp. 5052–5056, 1982.

[99] M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," *J. Appl. Phys.*, vol. 40, no. 1, pp. 278–283, 1969.

[100] M. Schulz and N. M. Johnson, "Evidence for multiphonon emission from interface states in MOS structures," *Solid State Commun.*, vol. 25, pp. 481–484, 1978.

[101] C. T. Sah, "Electronic processes and excess currents in gold-doped narrow silicon junctions," *Phys. Rev.*, vol. 123, no. 5, pp. 1594–1612, Sept. 1961.

[102] A. G. Chynoweth, W. L. Feldmann, and R. A. Logan, "Excess tunnel current in silicon esaki junctions," *Phys. Rev.*, vol. 121, no. 3, pp. 684–694, Feb. 1961.

[103] E. H. Nicollian and A. Goetzberger, "The Si-SiO_2 interface – electrical properties as determined by the metal-insulator-silicon conductance technique," *Bell Syst. Tech. J.*, vol. 46, pp. 1055–1133, 1967.

[104] E. H. Nicollian and J. R. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology.* John Wiley & Sons, 1982.

[105] N. Holonyak, I. A. Lesk, R. N. Hall, J. J. Tiemann, and H. Ehrenreich, "Direct observation of phonons during tunneling in narrow junction diodes," *Phys. Rev. Let.*, vol. 3, no. 4, pp. 167–168, Aug. 1959.

[106] M. Lax, "Giant traps," *J. Phys. Chem. Solids*, vol. 8, pp. 66–73, 1959.

[107] V. L. Bonch-Bruevich and E. G. Landsberg, "Recombination mechanisms," *Phys. Stat. Sol.*, vol. 29, no. 9, pp. 10–43, 1968.

[108] S. R. Dhariwal and P. T. Landsberg, "Electron-hole recombination via a simplified cascade process," *J. Phys.: Condens. Matter*, vol. 1, pp. 569–584, 1989.

[109] P. T. Landsberg and S. R. Dhariwal, "Electric field dependence of capture and emission rates by truncated cascade recombination," *Phys. Rev. B*, vol. 39, no. 1, pp. 91–93, Jan. 1989.

[110] D. V. Lang and C. H. Henry, "Nonradiative recombination at deep levels in GaAs and GaP by lattice-relaxation multiphonon emission," *Phys. Rev. Let.*, vol. 35, no. 22, Dec. 1975.

[111] H. Gummel and M. Lax, "Thermal capture of electrons in silicon," *Annals of Physics*, vol. 2, pp. 28–56, 1957.

[112] G. Rickayzen, "On the theory of the thermal capture of electrons in semiconductors," *Proc. Roy. Soc. A*, vol. 241, pp. 480–494, Sept. 1957.

[113] J. Bourgoin and M. Lannoo, *Point Defects in Semiconductors II: Experimental Aspects.* Springer-Verlag, 1983.

[114] J. Bardeen and W. Shockley, "Deformation potentials and mobilities in non-polar crystals," *Phys. Rev.*, vol. 80, no. 1, pp. 72–80, Oct. 1950.

[115] O. Madelung, *Introduction to Solid-State Theory.* Springer-Verlag, 1981.

[116] M. J. Kirton and M. J. Uren, "Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency $(1/f)$ noise," *Advances in Physics*, vol. 38, no. 4, pp. 367–468, 1989.

[117] M. J. Kirton and M. J. Uren, "Capture and emission kinetics of individual Si:SiO_2 interface states," *Appl. Phys. Lett.*, vol. 48, no. 19, pp. 1270–1272, 1986.

[118] B. J. Gross, *1/f Noise in MOSFETs with Ultrathin Gate Dielectrics.* PhD thesis, Massachusetts Institute of Technology, 1992.

[119] M. O. Andersson and O. Engström, "Atomic relaxation of si$-SiO_2$ interface states measured by a photo-depopulation technique," *Appl. Surface Sci.*, vol. 39, pp. 289–300, 1989.

[120] M. Aslam, "Electron self-trapping in SiO_2," *Appl. Phys. Lett.*, vol. 62, no. 1, pp. 159–162, 1987.

[121] M. O. Andersson, Z. Xiao, S. Norrman, and O. Engstrom, "Model based on trap-assisted tunneling for two-level current fluctuations in submicrometer metal–silicon-dioxide–silicon diodes," *Phys. Rev. B*, vol. 41, no. 14, pp. 9836–9842, 1990.

[122] F. Stern and W. E. Howard, "Properties of semiconductor surface inversion layers in the electric quantum limit," *Phys. Rev.*, vol. 163, no. 1, pp. 816–835, 1967.

[123] F. Stern, "Self-consistent results for n-type Si inversion layers," *Phys. Rev. B*, vol. 5, no. 12, pp. 4891–4899, 1972.

[124] Z. A. Weinberg, "Tunneling of electrons from Si into thermally grown SiO_2," *Solid-St. Electron*, vol. 20, pp. 11–18, 1977.

[125] W. E. Hagston and J. E. Lowther, "Multiphonon processes in rare-earth ions," *Physica*, vol. 70, pp. 40–61, 1973.

[126] J. M. Ziman, *Elements of Advanced Quantum Theory.* Cambridge University Press, 1969.

[127] N. W. Ashcroft and N. D. Mermin, *Solid State Physics.* Hold, Rinehart and Winston, 1976. Appendices L and 0.

[128] R. C. Jaklevic and J. Lambe, "Molecular vibration spectra by electron tunneling," *Phys. Rev. Let.*, vol. 17, no. 22, pp. 1139–1140, Nov. 1966.

[129] D. J. Scalapino and S. M. Marcus, "Theory of inelastic electron-molecule interactions in tunnel junctions," *Phys. Rev. Let.*, vol. 18, no. 12, pp. 459–461, March 1967.

[130] D. J. DiMaria, "Trap creation in silicon dioxide produced by hot electrons," *J. Appl. Phys.*, vol. 65, no. 6, pp. 2342–2356, 1989.

[131] E. O. Kane, "Theory of tunneling," *J. Appl. Phys.*, vol. 32, no. 1, pp. 83–91, Jan. 1961.

[132] C. T. Sah, "Excess currents in semiconductor tunneling," in *Tunneling Phenomena in Solids* (E. Burstein and S. Lundqvist, eds.), Plenum Press, New York, 1969.

[133] H. Overhof, "Hopping conductivity in disordered solids," *Advances in Solid State Physics*, vol. XVI, pp. 239–265, 1976.

[134] M. Pollak, "Temperature dependence of ac hopping conductivity," *Phys. Rev.*, vol. 138, no. 6A, pp. A1822–A1826, June 1965.

[135] J. Shewchun, A. Waxman, and G. Warfield, "Tunneling in MIS structures–I," *Solid-St. Electron*, vol. 10, pp. 1165–1186, 1967.

[136] E. O. Kane, "Zener tunneling in semiconductors," *J. Phys. Chem. Solids*, vol. 12, pp. 181–188, 1959.

[137] R. Stratton, "Volt-current characteristics for tunneling through insulating films," *J. Phys. Chem. Solids*, vol. 23, pp. 1177–1190, 1962.

[138] F. A. Padovani and R. Stratton, "Experimental energy-momentum relationship determination using schottky barriers," *Phys. Rev. Let.*, vol. 16, no. 26, pp. 1202–1204, June 1966.

[139] C. M. Chaves, N. Majlis, and M. Cardona, "Electronic surface states in germanium and silicon," *Solid State Communications*, vol. 4, pp. 271–274, 1966.

[140] G. Lewicki and C. A. Mead, "Experimental determination of E-k relationship in electron tunneling," *Phys. Rev. Let.*, vol. 16, no. 21, pp. 939–941, May 1966.

[141] J. Maserjian, "Tunneling in thin MOS structures," *J. Vac. Sci. Tech.*, vol. 11, no. 6, pp. 996–1003, Nov./Dec. 1974.

[142] J. Maserjian and G. P. Petersson, "Tunneling through thin MOS structures: Dependence on energy (E - κ)*," *Appl. Phys. Lett.*, vol. 25, no. 1, pp. 50–52, July 1974.

[143] G. P. Petersson, C. M. Svensson, and J. Maserjian, "Resonance effects observed at the onset of Fowler-Nordheim tunneling in thin MOS structures," *Solid-St. Electron*, vol. 18, pp. 449–451, 1975.

[144] G. Lewicki and J. Maserjian, "Oscillations in MOS tunneling," *J. Appl. Phys.*, vol. 46, no. 7, pp. 3032–3039, July 1975.

[145] W. Franz, "Dielektrischer durchschlag," in *Handbuch der Physik* (S. Flügge, ed.), pp. 201–219, Springer-Verlag, 1956.

[146] L. B. Freeman and W. E. Dahlke, "Theory of tunneling into interface states," *Solid-St. Electron*, vol. 13, pp. 1483–1503, 1970.

[147] Z. A. Weinberg, G. W. Rubloff, and E. Bassous, "Transmission, photoconductivity, and the experimental band gap of thermally grown SiO_2 films," *Phys. Rev. B*, vol. 19, no. 6, pp. 3107–3117, 1979.

[148] J. R. Chelikowsky and M. Schluter, "Electron states in α-quartz: A self-consistent pseudopotential calculation," *Phys. Rev. B*, vol. 15, no. 8, pp. 4020–4029, 1977.

[149] R. Williams and M. H. Woods, "High electric fields in silicon dioxide produced by corona charging," *J. Appl. Phys.*, vol. 44, no. 3, pp. 1026–1028, 1973.

[150] L. Esaki and P. J. Stiles, "New type of negative resistance in barrier tunneling," *Phys. Rev. Let.*, vol. 16, no. 24, pp. 1108–1111, June 1966.

[151] R. F. Pierret, *Advanced Semiconductor Fundamentals.* Modular Series on Solid State Devices, Addison-Wesley, 1987.

[152] K. S. Ralls, W. J. Skocpol, L. D. Jackel, R. E. Howard, L. A. Fetter, R. W. Epworth, and D. M. Tennant, "Discrete resistance switching in submicrometer silicon inversion layers: Individual interface traps and low frequency (1/f) noise," *Phys. Rev. Let.*, vol. 52, no. 3, p. 228, 1984.

[153] C. Surya and T. Y. Hsiang, "Surface mobility fluctuations in metal-oxide-semiconductor field-effect transistors," *Phys. Rev.*, vol. 35, no. 12, p. 6343, 1987.

[154] R. Jayaraman, *Reliability and 1/f Noise Properties of MOSFETs With Nitrided Oxide Gate Dielectrics.* PhD thesis, Massachesetts Institute of Technology, 1988.

[155] J. R. Brews, "Theory of the carrier-density fluctuations in an IGFET near threshold," *J. Appl. Phys.*, vol. 46, no. 5, p. 2181, 1975.

[156] C. T. Sah, T. H. Ning, and L. L. Tschopp, "The scattering of electrons by surface oxide charges and by lattice vibrations at the silicon–silicon dioxide interface," *Surface Science*, vol. 32, p. 561, 1972.

[157] N. Shiono, M. Shimaya, and O. Nakajima, "Charge character of interface traps at the SiO_2 interface," *Appl. Phys. Lett.*, vol. 48, no. 17, pp. 1129–1131, 1986.

[158] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, ch. 5, pp. 238–254. Dover, 1965.

[159] L. Lundkvist, I. Lundstrom, and C. Svensson, "Discharge of MNOS structures," *Solid-St. Electron*, vol. 16, pp. 811–823, 1973.

[160] K. Yamabe and Y. Miura, "Discharge of trapped electrons from MOS structures," *J. Appl. Phys.*, vol. 51, no. 12, pp. 6258–6264, 1980.

[161] V. Lakshmanna and A. S. Vengurlekar, "Logarithmic detrapping response for holes injected into SiO_2 and the influence of thermal activation and electric fields," *J. Appl. Phys.*, vol. 63, no. 9, pp. 4548–4554, 1988.

[162] P. M. Sutton, "Space charge and electrode polarization in glass, I," *J. Am. Ceram. Soc.*, vol. 47, no. 4, pp. 188–194, April 1964.

[163] K. Hyyppa, "Dielectric absorption in memory capacitors," *IEEE Trans. Instrum. Meas.*, pp. 53–56, Feb. 1972.

[164] J. N. Wright, "A method of compensating for dielectric absorption in capacitors," Master's thesis, MIT, 1978.

[165] J. W. Gray, "Integrator with means for compensating for capacity absorption effects," July 1962. U.S. Patent No. 3,047,808.

[166] J. W. Fattaruso, M. D. Wit, G. Warwar, K.-S. Tan, and R. K. Hester, "The effect of dielectric relaxation on charge-redistribution A/D converters," *IEEE J. Solid-State Circuits*, vol. 25, pp. 1550–1561, Dec. 1990.

[167] J. W. Fattaruso, M. de Wit, G. Warwar, K. S. Tan, and R. K. Hester, "The effect of dielectric relaxation on charge-redistribution A/D converters," in *Symposium on VLSI Circuits*, pp. 29–30, 1990.

[168] P. C. Dow, "An analysis of certain errors in electronic differential analyzers II – capacitor dielectric absorption," *IRE Trans. Electronic Computers*, pp. 17–22, March 1958.

[169] P. M. Sutton, "Space charge and electrode polarization in glass, II," *J. Am. Ceram. Soc.*, vol. 47, no. 5, pp. 219–230, May 1964.

[170] H. E. Taylor, "An investigation of the dielectric relaxation exhibited by glass. Part II. Soda-silica and soda-lime-silica glasses," *J. Soc. Glass Technol.*, vol. 43, pp. 124T–146T, 1959.

[171] A. H. M. Shousha and L. Young, "Space charge contribution to the polarization current in thin amorphous dielectric films," *Thin Solid Films*, vol. 8, pp. 383–396, 1971.

[172] R. Dreiner, "Charge storage effects in tantalum oxide films," *J. Electrochem. Soc.*, vol. 111, no. 1, pp. 27–34, Jan. 1964.

[173] F. Cardon, "Relation between polarization phenomenon and space-charge-limited currents in rutile single crystals (TiO_2)," *J. Appl. Phys.*, vol. 33, no. 11, pp. 3358–3365, Nov. 1962.

[174] C. Cherki, R. Coelho, and R. Nannoni, "Influence of interfacial tunnel exchange on dielectric losses in thin amorphous insulating films," *Phys. Stat. Sol. (A)*, vol. 2, pp. 785–796, 1970.

[175] G. Caserta, B. Rispoli, and A. Serra, "Space-charge-limited current and band structure in amorphous organic films," *Phys. Stat. Sol.*, vol. 35, pp. 237–248, 1969.

[176] G. Lengyel, "Schottky emission and conduction in some organic insulating materials," *J. Appl. Phys.*, vol. 37, no. 2, pp. 807–810, Feb. 1966.

[177] W. A. Phillips, "Low temperature dielectric relaxation in polyethylene and related hydrocarbon polymers," *Proc. Roy. Soc. Lond. A.*, vol. 319, pp. 565–581, 1970.

[178] H. J. Wintle, "Absorption currents and steady currents in polymer dielectrics," *J. Non-Crystalline Solids*, vol. 15, pp. 471–486, 1974.

[179] D. K. Das Gupta and K. Joyner, "On the nature of absorption currents in polyethylene terephthalate (PET)," *J. Phys. D: Appl. Phys.*, vol. 9, pp. 829–840, 1976.

[180] D. M. Taylor and T. J. Lewis, "Electrical conduction in polyethylene terephthalate and polyethylene films," *J. Phys. D: Appl. Phys.*, vol. 4, pp. 1346–1357, 1971.

[181] J. Lindmayer, "Current transients in insulators," *J. Appl. Phys.*, vol. 36, no. 1, pp. 196–201, Jan. 1965.

[182] P. J. Burkhardt, "Dielectric relaxation in thermally grown SiO_2 films," *IEEE Trans. Electron Devices*, vol. ED-13, no. 12, pp. 268–275, Feb. 1966.

[183] A. K. Jonscher, *Dielectric Relaxation in Solids.* Chelsea Dielectrics Press, 1983.

[184] V. V. Daniel, *Dielectric Relaxation.* Academic Press, 1967.

[185] K. S. Cole and R. H. Cole, "Dispersion and absorption in dielectrics – I. alternating current characteristics," *J. Chem. Phys.*, vol. 9, April 1941.

[186] K. S. Cole and R. H. Cole, "Dispersion and absorption in dielectrics – II. direct current characteristics," *J. Chem. Phys.*, vol. 10, Feb. 1964.

[187] J. M. Stevels and J. Volger, "Impurity-induced imperfections and the dielectric properties of quartz crystals," *Phys. Chem. Glasses*, vol. 4, no. 6, pp. 247–252, 1963.

[188] E. H. Snow and B. E. Deal, "Polarization effects in insulating films on silicon—a review," *Trans. Met. Soc. AIME*, vol. 242, pp. 512–523, March 1968.

[189] R. H. Walden, "A method for the determination of high-field conduction laws in insulating films in the presence of charge trapping," *J. Appl. Phys.*, vol. 43, no. 3, pp. 1178–1186, March 1972.

[190] J. G. Simmons and G. W. Taylor, "Dielectric–relaxation currents in insulators," *Phys. Rev. B*, vol. 5, no. 2, pp. 553–556, Jan. 1972.

[191] H. E. Taylor, "An investigation of the dielectric relaxation exhibited by glass. Part I. Theory and methods of measurement," *J. Soc. Glass Technol.*, vol. 41, pp. 350T–382T, 1957.

[192] J. R. Macdonald, "Electrical response of materials containing space charge with discharge at the electrodes," *J. Chem. Phys.*, vol. 54, no. 5, pp. 2026–2050, March 1971.

[193] A. K. Jonscher, "Frequency-dependence of conductivity in hopping systems," *J. Non-Crystalline Solids*, vol. 8-10, pp. 293–315, 1972.

[194] A. K. Jonscher, "Hopping losses in polarisable dielectric media," *Nature*, vol. 250, pp. 191–193, 1974.

[195] A. K. Jonscher, F. Meca, and H. M. Millany, "Charge-carrier contributions to dielectric loss," *J. Phys. C: Solid St. Phys.*, vol. 12, pp. L293–L296, 1979.

[196] P. Wilcox, "A dielectric loss model based on interfacial electron tunneling," *Canadian J. Phys.*, vol. 50, pp. 912–924, 1972.

[197] H. J. Wintle, "Absorption current, dielectric constant, and dielectric loss by the tunneling mechanism," *J. Appl. Phys.*, vol. 44, no. 6, pp. 2514–2519, June 1973.

[198] W. H. Beyer, ed., *CRC Standard Mathematical Tables*, p. 399. CRC Press, 26 ed., 1981.

[199] W. H. Beyer, ed., *CRC Standard Mathematical Tables*, p. 343. CRC Press, 26 ed., 1981.

[200] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, p. 310. Academic Press, 1980.

[201] D. A. Buchanan and D. J. Dimaria, "Interface and bulk trap generation in metal-oxide-semiconductor capacitors," Research Report RC 14815 (#66376), IBM, March 1989.

[202] J. K. Roberge, *Operational Amplifiers: Theory and Practice.* John Wiley & Sons, 1975.

[203] S. K. Lai, J. Lee, and V. K. Dham, "Electrical properties of nitrided-oxide systems for use in gate dielectrics and EPROM," *IEDM Technical Digest*, p. 190, 1983.

[204] F. L. Terry, M. L. Naiman, R. J. Aucoin, and S. D. Senturia, "Megarad-resistant 10nm gate dielectrics," *IEEE Trans. Nucl. Sci.*, vol. NS-28, no. 6, p. 4389, 1981.

[205] F. L. Terry, R. J. Aucoin, M. L. Naiman, and S. D. Senturia, "Radiation effects in nitided oxides," *IEEE Electron Device Lett.*, vol. EDL-4, p. 191, 1983.

[206] S. S. Wong, C. G. Sodini, T. W. Ekstedt, H. R. Grinolds, K. H. Jackson, and S. H. Kwan, "Low pressure nitrided-oxide as a thin gate dielectric for MOSFET's," *J. Electrochem. Soc.*, vol. 130, p. 1139, 1983.

[207] T. Ito, T. Nakamura, and H. Ishikawa, "Effect of thermally nitrided SiO_2 (nitroxide) on MOS characteristics," *J. Electrochem. Soc.*, vol. 129, no. 1, p. 184, 1982.

[208] C. T. Chen, F. C. Tseng, C. Y. Chang, and M. K. Lee, "Study of electrical characteristics on thermally nitrided SiO_2 (nitroxide) films," *J. Electrochem. Soc.*, vol. 131, p. 875, 1984.

[209] F. L. Terry, *Electrical Properties of Nitrided Silicon Dioxide as an MOS Gate Insulator.* PhD thesis, Massachesetts Institute of Technology, 1985.

[210] B. J. Gross, K. S. Krisch, and C. G. Sodini, "An optimized 850°C low-pressure-furnace reoxidized nitrided oxide (ROXNOX) process," *IEEE Trans. Electron Devices*, vol. 38, no. 9, pp. 2036–2041, Sept. 1991.

[211] S. Selberherr, *Analysis and Simulation of Semiconductor Devices.* Springer-Verlag, 1984.

[212] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, pp. 916–919. National Bureau of Standards, 10 ed., 1972.

[213] W. H. Press, B. P. Flannery, S. A. Teulolsky, and W. T. Vetterline, *Numerical Recipes in C.* Cambridge University Press, 1990.

[214] R. Jayaraman and C. G. Sodini, "A 1/f noise technique to extract the oxide trap density near the conduction band edge of silicon," *IEEE Trans. Electron Devices*, vol. 36, pp. 1773–1782, 1989.

[215] A. Papoulis, *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, 1984.

[216] Z. Celik-Butler and T. Y. Hsiang, "Spectral dependence of $1/f^{\gamma}$ noise on gate bias in n-MOSFETs," *Solid-St. Electron*, vol. 30, no. 4, pp. 419–423, 1987.

[217] P. K. Chakraborty and J. C. Biswas, "A study of excess current in an Esaki junction," *J. Appl. Phys.*, vol. 64, no. 11, Dec. 1988.

[218] J. Price, "Tunneling from trap states in Esaki diodes," *Bull. Amer. Phys. Soc.*, vol. 5, p. 406, 1960.

[219] C. B. Duke and M. E. Alferieff, "Field emission through atoms adsorbed on a metal surface," *J. Chem. Phys.*, vol. 46, no. 3, Feb. 1967.

[220] S. Chaudhuri, D. D. Coon, and R. P. G. Karunasiri, "Impurity-to-band tunneling in semiconductors," *J. Appl. Phys.*, vol. 54, no. 9, Sept. 1983.

[221] W. W. Anderson and H. J. Hoffman, "Field ionization of deep levels in semiconductors with applications to $Hg_{1-x}Cd_xTe$ $p-n$ junctions," *J. Appl. Phys.*, vol. 53, no. 12, Dec. 1982.

[222] E. N. Korol, "Ionization of impurity states in semiconductors by an electric field," *Sov. Phys. Solid State*, vol. 19, no. 8, pp. 1327–1330, Aug. 1977.

[223] S. Chaudhuri, D. D. Coon, and G. E. Derkis, "Quantum-mechanical estimates of the speed of field ionization of shallow impurity levels," *Appl. Phys. Lett.*, vol. 37, no. 1, pp. 111–113, July 1980.

[224] J. R. Banavar, D. D. Coon, and G. E. Derkis, "Low-temperature field ionization of localized impurity levels in semiconductors," *Appl. Phys. Lett.*, vol. 34, no. 1, pp. 94–96, Jan. 1979.

[225] S. Chaudhuri, D. D. Coon, G. E. Derkits, and J. R. Banavar, "Rate of field ionization from S states with a quantum defect," *Phys. Rev. A*, vol. 23, no. 4, pp. 1657–1661, April 1981.

[226] J. R. Oppenheimer, "Three notes on the quantum theory of aperiodic effects," *Phys. Rev.*, vol. 13, pp. 66–81, Jan. 1928.

[227] J. Bardeen, "Tunneling from a many-particle point of view," *Phys. Rev. Let.*, vol. 6, no. 2, pp. 57–59, Jan. 1961.

[228] P. K. Hansma, ed., *Tunneling Spectroscopy – Capabilities, Applications, and New Techniques.* Plenum Press, 1982.

[229] Baym, *Lectures on Quantum Mechanics.* The Benjamin/Cummings Publishing Co., 1969.

[230] K. Hess, *Advanced Theory of Semiconductor Devices*, ch. Appendix A, pp. 238–241. Prentice Hall, 1988.

[231] C. B. Duke, *Tunneling in Solids.* Solid State Physics, Supplement 10, Academic Press, 1969.

[232] Z. A. Weinberg and A. Harstein, "Photon assisted tunneling from aluminum into silicon dioxide," *Solid State Commun.*, vol. 20, pp. 179–182, 1976.

[233] R. L. Liboff, *Introductory Quantum Mechanics.* Holden-Day, Inc., 1980.

[234] I. Lündstrom and C. Svensson, "Tunneling to traps in insulators," *J. Appl. Phys.*, vol. 45, no. 12, pp. 5045–47, Dec. 1972.

[235] D. J. BenDaniel and C. B. Duke, "Space-charge effects on electron tunneling," *Phys. Rev.*, vol. 152, no. 2, pp. 152–692, Dec. 1966.

[236] E. Merzbacher, *Quantum Mechanics.* John Wiley and Sons, 1970. second edition.

[237] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics*, vol. 3. Pergamon Press, 1977. Problem 3.

[238] J. W. Corbett, "Response of anisotropic defects in solids to stress," *Am. J. Phys.*, vol. 31, pp. 521–536, 1963.

[239] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, vol. III. Addison-Wesley, 1965.

[240] V. Narayanamurti and R. O. Pohl, "Tunneling states of defects in solids," *Reviews of Modern Physics*, vol. 42, no. 2, pp. 201–236, April 1970.

[241] J. A. Sussman, "Phonon induced tunneling of ions in solids," *Phys. kondens. Materie*, vol. 2, pp. 146–160, 1964.

[242] P. W. Anderson, B. I. Halperin, and C. M. Varma, "Anomolous low-temperature thermal properties of glasses and spin glasses," *The Philosophical Magazine*, vol. 25, pp. 1–9, 1971.

[243] W. A. Phillips, "Tunneling states in amorphous solids," *Journal of Low Temperature Physics*, vol. 7, no. 3/4, pp. 351–360, 1972.

[244] J. Jackle, "On the ultrasonic attenuation in glasses at low temperatures," *Z. Physik*, vol. 257, pp. 212–223, 1972.

[245] S. Gasiorowicz, *Quantum Physics.* John Wiley and Sons, 1974.

[246] J. S. Blakemore, "Approximations for Fermi-Dirac integrals, especially the function $\mathcal{F}_{1/2}(\eta)$ used to describe electron density in a semiconductor," *Solid-St. Electron*, vol. 25, no. 11, pp. 1067–1076, 1982.