CALT 68-xxxx
CNS Thesis

Foundations of Learning in Analog VLSI

Thesis by

Paul Hasler

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1997
(Submitted February 3, 1997)

© 1997

Paul Hasler

All Rights Reserved

Acknowledgements

I wish to acknowledge several people who supported, listened, advised or assisted in my thesis work:

My wife Beth for standing by me and making the last four years memorable,

My parents for their constant support and encouragement,

Carver Mead for his support, guidance, inspiration, and technical insight, as well as showing me by example how ideally to run a laboratory,

Candace Schmidtt, Donna Fox, Calvin Jackson, and Jim Campbell for providing such an organized and efficient place to work, and for taking care of so many 'small' details that add up over the course of five years,

Brad Minch: it is a rare blessing to work not only with a valued colleague, but a good friend who shares a common faith,

Andreas Andreou for his friendship, his valued comments, and for helping me formulate Chapter 2 into a clean story,

Chris Diorio: our collaboration pushed each of us to do better and greater work,

Rahul Sarpeshkar for several discussions on noise, dynamic range, cochleas, device physics,

The rest of my thesis committee, Christof Koch, Demetri Psaltis, and Rod Goodman,

Past and present Caltech students, Ron Benson, Buster Boahen, Tobi Delbruck, Tim Horiuchi, Fritz Kruger, Sanjoy Mahajan, Lena Peterson, Shih-Chii Lui, Theron Stanford, Lloyd Watts, and Orly Yadid-Pecht,

The 95'-96' Analog VLSI and Neural Systems class of students for showing me the failings of many floating-gate circuits, and for creatively breaking many of these circuits,

Lyn Dupre for her patience in editing this manuscript and several related papers,

The Office of Naval Research, the Advanced Research Projects Agency, the Beckman Hearing Institute, and the NSF Center for Neuromorphic Systems Engineering for their support of this work.

Abstract

Floating-gate technology can be used to build silicon systems that adapt and learn. This technology is well suited to implement adaptation and learning because we are not building analog EEPROMS, but rather circuit elements with important time-domain dynamics. These floating-gate circuits use the hot-electron-injection, electron-tunneling, and drain-induced-barrier-lowering phenomena in a standard submicron CMOS process. This technology works with the constraints of the silicon medium, and is similar to biological systems that turned potential liabilities into features.

I develop the first analytical model of the impact-ionization and hot-electron processes in MOS devices by solving for a self-consistent distribution function from the spatially varying Boltzmann transport equation. From this electron distribution function, the probabilities of impact ionization and hot-electron injection are calculated as functions of channel current, drain voltage, and floating-gate voltage. The analytical model simultaneously fits both the hot-electron-injection and impact-ionization data. These analytical results yield measurements of the energy-dependent impact-ionization collision rate that is consistent with numerically calculated collision rates reported in the literature.

I describe the design, fabrication, characterization, and modeling of an array of single-transistor synapses that simultaneously store the weight value, compute the product of the input and floating gate value, and update the weight value according to a hebbian or backpropagation learning rule. Circuits with one floating-gate synapse exhibit a range of possible stabilizing and destabilizing behaviors, and circuits with multiple-synapses show examples of competitive and cooperative behavior. By providing feedback to the source, we get a *p*FET synapse where voltage changes in both the floating gate and drain stabilize the floating gate.

I present a bandpass floating-gate amplifier that uses tunneling and *p*FET hot-electron injection to adaptively set its DC operating point. Because the gate cur-

rents are small, the circuit exhibits a high-pass characteristic with a cutoff frequency less than 1 Hz. The high frequency cutoff is controlled electronically, as is done in continuous-time filters. I have derived analytical models that completely characterize the amplifier and that are in good agreement with experimental data for a wide range of operating conditions and input waveforms. This autozeroing floating-gate amplifier demonstrates how to use continuous-time, floating-gate adaptation.

Contents

Acknowledgements iii

Abstract iv

1 Floating-Gate Learning Systems 1

1.1 Historical Perspective on Analog Implementations of Neural Systems 2

1.2 Floating-Gate Technology 6

1.3 Thesis Overview: Electron Transport to Floating-Gate Circuits 9

2 Impact Ionization and Hot-Electron Injection in MOSFETs Derived Consistently from Boltzman Transport 12

2.1 Electron Transport in the Drain to Channel Depletion Region 16

2.1.1 Boltzmann Transport Under an Applied Electric Field 16

2.1.2 Simplifications for Hot-Electron Transport in the MOSFET Drain-to-Channel Region 17

2.1.3 Collision Operators 19

2.2 Solving the Electron Distribution Function 24

2.2.1 Average Electron Energy and Momentum Direction 25

2.2.2 Solution of the Distribution function 28

2.3 Comparing Theory with Experiment 33

2.3.1 Device Structure and Bias Condition 34

2.3.2 Drain-to-Channel Depletion Width Dependence on Φ_{dc} 35

2.3.3 Electron Distribution Function at the Drain Edge 37

2.3.4 Modeling of Hot-Electron Injection in MOS Transistor 41

2.3.5 Electron Impact Ionization 47

2.4 Discussion 51

2.4.1 Effect of Collision Broadening on the Phonon Collision Operator 51
2.4.2 Effect of Momentum Scattering Optical Phonons 53
2.4.3 How the Electron Gets the Necessary Angle for Impact Ionization 57

3 Single-Transistor Synapses 59
3.1 Overview of Single-Transistor Learning Synapses 59
3.2 Nonadaptive Behavior 62
3.3 Electron Tunneling 65
3.3.1 Hot-Electron Injection 68
3.4 An Array of Single Transistor Synapses 74

4 Continuous-Time Feedback in Floating-Gate MOS Circuits 79
4.1 The Source-Degenerated *p*FET Synapse 81
4.2 Stability of Single-Synapse Circuits with Floating-Gate Feedback . . . 86
4.2.1 *p*FET Floating-Gate Circuits 88
4.2.2 *n*FET Floating-Gate Circuits 92
4.2.3 Source-Degenerated *p*FET Floating-Gate Circuits 94
4.3 Networks of Two Coupled Synapses 97
4.4 Conclusions 103

5 An Autozeroing Floating-Gate Amplifier 105
5.1 Qualitative Presentation of AFGA Operation 107
5.2 Equilibrium Voltages of the AFGA 110
5.3 Low-Frequency AFGA Behavior 114
5.3.1 Low-Frequency Model 114
5.3.2 Response to a Voltage Step 117
5.3.3 Long-Term Parameter Drift 118
5.4 High-Frequency AFGA Behavior 120
5.5 Frequency Response of the AFGA 124
5.6 Steady-State Output-Voltage Dependence on the Output Signal . . . 130
5.7 Other AFGA Effects 133

5.7.1 Model of the Above-Threshold AFGA 133
5.7.2 Continuous Operation of the Tunneling Current 135
5.7.3 Restoration to Equilibrium when the Output Voltage Starts at a Supply Rail 137
5.8 Conclusions 137

6 Conclusions and Future Directions 139
6.1 Summary of Thesis Results 139
6.2 Future Directions 142
6.2.1 Autozeroing Second-Order Section 142
6.2.2 Adaptive Winner-Take-All 144

Bibliography 147

A Appendix to Chapter 2 157
A.1 Derivation of Electron Characteristics 157
A.2 Distribution Function with Impact Ionization Losses 158
A.2.1 Derivation of the $a(z, E)$ Equation 158
A.2.2 Simplification of $a(z, E)$ under Electric Fields 159

List of Figures

1.1 In previous neural network implementations, the adaptive elements (synapses) were complex and required a tremendous amount of area. (a) Synapse layout from [5]; the size of this synapse was typical of dense synapse designs of similar complexity. (b) Typical implementations used separate computation, memory, and adaptation blocks, because these networks were direct implementations of mathematical models. 4

1.2 Transistor leakage currents limit adaptation time-constants. (a) Circuit schematic when storing a voltage on a capacitor. (b) Plot of the resulting stored voltage versus time. The constant leakage current decreases linearly the stored voltage over time. 5

1.3 Illustration of our floating-gate CMOS technology. (a) Charge on the floating-gate is nearly permanently stored because electrons are surrounded by a high quality insulator. (b) Since the floating-gate voltage can modulate a MOSFET's channel current, the floating gate is not only a memory device, but also can be an integral part of a computation. (c) The floating-gate charge can be modified by electron tunneling and hot-electron injection. Continuous-time adaptation is possible if the floating-gate current is a function of other device parameters. In this respect, hot-electron injection is an important adaptation mechanism, because floating-gate current is proportional to channel current and an exponential function of drain voltage. . 11

2.1 Cross section of the MOSFET device we used to measure the hot-electron effects. It uses a moderately doped ($1\text{x}10^{17}\text{cm}^{-3}$) substrate to achieve a high threshold voltage which allows hot-electron injection for bias current levels in subthreshold. The higher doping is consistent with a 0.2μm channel length CMOS process; therefore the effects are directly applicable to modern processes although the device was fabricated in a 2μm process. The $n$well isolates the moderately doped substrate region from the surrounding substrate, and allows measurement of substrate current. Holes resulting from impact ionization are measured at the $p$base contact. The hot-electron injection process is identical for the FET with or without the isolating $n$well. Inset: the electron is accelerated through the drain depletion region (1), and when it gains energy greater than the Si–SiO_2 barrier, the electron is injected over the Si–SiO_2 barrier to the floating-gate (2). Lower Left: Circuit symbol for the nFET with the $p$base implant. 14

2.2 Graphical representations of the phonon collision operator. An electron can either gain or lose energy due to a phonon. The transition probabilities of gaining or losing energy due to a phonon are different. The operator is based on a mean free path of a collision; therefore the longer an electron goes in a region, the more likely it will gain or lose energy due to a phonon collision. Gaining or losing a phonon will alter the electron's momentum as well as its energy. 20

2.3 A second way to graphically represent the phonon collision operator is considering an electron either gaining or losing energy due to phonons with equal strength along a particular coordinate system. 21

2.4 Plot of previous calculations of impact-ionization rate versus electron energy in silicon, and our derived impact-ionization rate from our measured impact-ionization and hot-electron injection data. I have assumed a constant velocity, since our model measures the impact-ionization mean-free length. Our measured data is directly related to $L(E)$ and not τ_{ion}. . . . 22

2.5 Band diagram illustrating hot-electron injection in a MOSFET biased in subthreshold. The appropriate variables in the Boltzmann transport equation and its variable transformations are shown on the graphs. (a) Band diagram along the surface of the Si-SiO_2 barrier. This region is the lowest local potential in either material; therefore the electrons are most likely to travel along this path. This region corresponds to path (1) in the inset in Fig. 2.1. (b) Band diagram at the drain edge from the substrate to the gate. This region corresponds to path (2) in the inset in Fig. 2.1. 24

2.6 Illustration of the electron characteristic paths at different electric fields. In the warm field region, the restoring force due to collisions is larger than the force due to the electric field. In the high field region, the force due to the electric field is larger than the restoring force due to collisions. When an electron reaches an electric field larger than the restoring force due to collisions it can gain significant energies above the conduction band. . . 27

2.7 Picture of the distribution function for an electron in the drain-to-channel beyond $z = z_{crit}$. This figure compares my Gaussian approximation model versus the solution using the exact boundary condition. For large positive energies, the distribution function does not change as fast as the Gaussian, but rather at a slope around kT_c. 30

2.8 Plot of $f(z, E)$ as a function of energy (E) for several locations in the drain-to-channel depletion region for $\Phi_{dc} = 4.0$V, λ =6.5nm, and $N_a = 10^{17}$cm^{-3}. For these parameters, $z_{crit} = 0.36d$. I normalized $f(z, E)$ over energy; normalizing over position would simply change the global scale factor. The width of the drain-to-channel depletion region is d. 32

2.9 Measured Early voltage versus Φ_{dc} for two values of gate voltage. This log-log plot brings out the power law of the junction, and therefore the effective doping profile. For a particular gate voltage, the Early voltage curve is consistent with results from a step junction with $L = 16\mu$m and $N_a = 9.53 \times 10^{16}$cm^{-3}. The curves are similar to the lower V_g case up to gate voltages of 7V, and then gradually increases until it reaches the higher V_g curve. For a gate voltage of 9.72eV, the Early voltage curve is consistent with results from a step junction with $N_a = 2 \times 10^{18}$cm^{-3}. 37

2.10 Computed parameters for the drain-to-channel depletion region. The plots show the width of the drain-to-channel depletion width, and the maximum electric field in the depletion region (at the drain edge). The substrate doping was 1×10^{17}cm^{-3}, and $\lambda = 6.5$nm. For this device, z_{crit} is approximately one third of the distance into the depletion region for $\Phi_{dc} = 3.6$V. 38

2.11 Plot of the average electron energy as a function for the drain-to-channel depletion potential. The substrate doping was 1×10^{17}cm^{-3}, and $\lambda = 6.5$nm. 39

2.12 Plot of z_{crit} versus MOSFET substrate doping for $\lambda = 6.5$nm. 40

2.13 Plot of the average electron energy at the drain edge versus drain-to-channel potential for four different MOSFET substrate dopings. 41

2.14 Attenuation function at the drain edge due to impact ionization collisions as a function of electron energy for five different Φ_{dc}. I calculated these curves by evaluating (2.56) by numerical integration. 42

2.15 1 - $a(z, E)$ at the drain edge due to impact ionization collisions as a function of electron energy for three different Φ_{dc}. I calculated these curves by evaluating (2.56) by numerical integration. 43

2.16 Plot of the distribution function at the drain edge as a function of energy for three drain-to-channel potentials. I calculated these curves by evaluating (2.56) by numerical integration. 44

2.17 Electron injection efficiency versus source current for three values of drain voltage. The curves were obtained by holding the drain and floating-gate fixed while sweeping the source voltage. Electron injection efficiency is defined as the ratio of the injection current and the source current. For subthreshold currents, this ratio is constant for a given drain voltage since the drain-to-channel voltage is constant in this regime. 45

2.18 (a) Measurements of hot-electron-injection efficiency versus drain-to-channel voltage for several values of source current. I simultaneously measured the substrate and gate currents for different drain-to-channel voltages which gives us the impact-ionization and hot-electron-injection efficiency. The drain-to-channel voltage is computed from the source current and the drain-to-source voltage. For each sweep, I used a constant gate voltage to choose a particular channel current; the actual oxide barrier height changes slightly due to image force lowering, because the floating-gate–to–drain voltage is not constant. E_{ox} will roughly change by 100meV as predicted by image force lowering. (b) Measurement of hot-electron-injection efficiency versus Φ_{dc} compared with a curve fit to the analytic model in (2.61). 46

2.19 Measurements of impact-ionization efficiency vs. drain to channel voltage for three source currents (gate voltages). 47

2.20 Measurement of impact-ionization efficiency versus Φ_{dc} compared with a curve fit to the analytic model in (2.71). The source voltage was held at zero and the gate voltage was held at 5.8V. 50

2.21 Illustration of the 2D nature of the electron transport in the drain-to-channel region. The electron elastically bumps along the barrier at the SiO_2 interface and then gets redirected back towards the interface due to the electric field. In this way, some of the electrons have the proper direction to enter the SiO_2 and eventually reach the floating gate. 57

3.1 Cross section of the nFET and pFET single-transistor synapses in an $n$well MOSIS process. The tunneling junctions used by the single-transistor synapses is a region of gate oxide between the polysilicon floating-gate and $n$well. For the nFET synapse, the pbase implant results in a larger threshold voltage, which results in all the electrons reaching the top of the SiO_2 barrier being swept into the floating gate. The pFET transistor is the standard pFET transistor in the $n$well process. 60

3.2 Circuit symbols of the nFET and pFET single-transistor synapses. (a) Circuit diagram of the nFET single-transistor synapse with its source connected to ground. (b) Layout of the nFET single-transistor synapse. (c) Circuit diagram of the pFET single-transistor synapse with its source connected to V_{dd}. (d) Layout of the pFET single-transistor synapse. 61

3.3 Tunneling in an $n$well process. (a) The tunneling junction is the capacitor between the floating gate and the $n$well; we use high-quality gate oxide to reduce the effects of electron trapping. Over a wide range of oxide voltage, most of the tunneling occurs between the floating gate and n^+ diffusion region because this region is accumulated and the higher electric fields at the corner of the floating gate. (b) Band diagram through the tunneling capacitor. For sufficiently large applied electric fields, the electron barrier becomes thin enough that an electron might tunnel through the barrier. 64

3.4 Electron tunneling current versus 1/oxide voltage. The two straight line fits are to the classic Fowler-Nordheim expression in (3.8). The two different straight-line regions might be due to tunneling through intermediate traps, or due to initially tunneling through the junction edge for low oxide voltages and tunneling through the middle of the junction for high oxide voltages. 66

3.5 Electron tunneling current at a fixed oxide voltage ($V_{ox} = 33$V) versus charge through our tunneling junction. This experiment induced 60V of charge on a 200pF capacitor; to first order, tunneling trap creation is a function of the total charge through the oxide. This measurement characterizes the electron-trapping effect in our tunneling-junction oxide; the degree that the tunneling current decreases shows the effect of increasing electron traps in the oxide. 67

3.6 Band diagram of a subthreshold pFET transistor under conditions favorable for hot-electron injection. The source of electrons to be injected to the floating-gate is created by hole impact ionization. E_{ox} is the Si–SiO_2 barrier, which is 3.04eV for no field across the oxide. 68

3.7 Measured data of the current dependences in hot-electron injecting pFETs versus the drain-to-channel voltage for two source currents. The top plot shows the ratio of impact-ionization current and source current versus drain-to-channel potential. The lower plot shows the ratio of hot-electron injection and impact-ionization current versus drain-to-channel potential. This ratio saturates around $\Phi_{dc} = 9.0$V due to electrons gaining sufficient energy to surmount the SiO_2 barrier, and gives evidence that hole-impact ionization is the source of the electrons. 69

3.8 nFET and pFET hot-electron injection efficiency ($\frac{I_{inj}}{I_s}$) versus Φ_{dc} for two values of source current. Injection efficiency is the ratio of injection current to source current. The two different source current values are nearly equal, which is consistent with injection efficiency being independent of source current. I show the linearized slope (V_{inj}) on this exponential scale for two Φ_{dc} biases. The slope of both curves on this exponential scale decrease with increasing Φ_{dc}. The moderately doped nFET substrate ($1 \times 10^{17}\text{cm}^{-3}$) increases the efficiency of the nFET hot-electron–injection process by increasing the electric field in the channel. 71

3.9 Plot of $\frac{d}{dt}\log(I_s)$ as a function of I_s for three different tunneling and three different drain voltages. Starting at an appropriately low (or high) source current (I_s), I measured this data by stepping to the desired drain or tunneling voltage, and letting the tunneling (or injection) current decrease (or increase) the synapse current. The drain voltage was 2V when tunneling, and the tunneling voltage was 23V when injecting. This measurement gives the floating-gate weight update rule as a function of I_s; plotting the data in this way shows the power-law dependence of channel current on the floating-gate current. 72

3.10 Plot of $\frac{d}{dt}\log(I_s)$ as a function of I_s for four different tunneling and three different drain voltages. Plotting the data in this way shows the power-law dependence of channel current on the floating-gate current. I measured this injection (tunneling) data as in Fig. 3.9 by starting the synapse at a low (or high) source current. Since this is a pFET synapse, the tunneling voltages are referenced to V_{dd}. 73

3.11 Circuit diagram of the single-transistor synapse array. Each transistor has a floating gate capacitively coupled to an input column line. A 2 x 2 section of the array allows us to characterize how modifying a single floating gate (such as synapse (1,1)) affects the neighboring floating gate values. The synapse currents are a measure of the synaptic weights, and are summed along each row by the source (V_s) or drain (V_d) lines into a typical soma circuit. 75

3.12 Output currents from a 2 x 2 section of the synapse array, showing 180 injection operations followed by 160 tunneling operations. Because our measurements from the 2 x 2 section come from a larger array, we also display the 'background' current from all other synapses on the row. This background current is several orders of magnitude smaller than the selected synapse current, and therefore negligible. 76

3.13 (a) Synapse (1,1) source current increment versus source current for several values of tunneling voltage. (b) Source current decrement during injection versus source current for several values of drain voltage. 77

4.1 (a) Circuit diagram and small-signal model of the nFET single-transistor synapse with its source connected to ground. The small-signal model assumes a constant tunneling current and that the parameters are positive. (b) Circuit diagram and small-signal model of the pFET single-transistor synapse with its source connected to V_{dd}. The small-signal model assumes a constant tunneling current and that the parameters are positive. The small-signal resistance from floating-gate to ground is negative due to the hot-electron injection currents. 80

4.2 The source-degenerated pFET single-transistor synapse. (a) Circuit of the source-degenerated pFET synapse. (b) The small-signal of the source-degenerated pFET synapse, where I have defined all of the small signal quantities to be positive. From this circuit, we can see that this pFET's floating-gate currents provide stabilizing feedback to the floating-gate and drain voltages. 82

4.3 Plot of source-degenerated pFET drain current versus gate voltage for three values of κ_x. Decreasing κ_x decreases the transistor's effective gate coupling to the surface potential (due to the decreasing exponential slope). 83

4.4 Plot of $\frac{d\log(I_s)}{dt}$ versus source current (I_s) for four different values of drain voltage in the source-degenerated pFET. The transistor has stabilizing behavior for some regions of this graph; these portions are reflected in the small signal model. The synapse goes from unstable feedback to stable feedback due to the change in the bias Φ_{dc} of the synapse; larger Φ_{dc} requires a lower κ_x for stabilizing behavior. 84

4.5 The effect of κ_x on tunneling current in the source-degenerated pFET. Plot of $\frac{d}{dt}\log(I_s)$, which is proportional to the tunneling current, as a function of I_s for three different values of κ_x. The curves are nearly straight lines, and the slope of the curves increases for a decreasing κ_x. 85

4.6 The effect of κ_x on hot-electron-injection current in the source-degenerated pFET. Plot of $\frac{d}{dt}\log(I_s)$, which is proportional to the injection current, as a function of I_s for three different values of κ_x, including $\kappa_x = 0$. 86

4.7 Source current versus drain voltage for two short-channel pFETs in a 2μm process with the gate and source voltages at V_{dd}. The drain voltage is measured relative to V_{dd}. The DIBL results in an exponentially increasing current for subthreshold biases; for the 1.5μm pFET, the current increases an e-fold for a 276.1mV change in the drain voltage, and for the 1.75μm pFET, the current increases an e-fold for a 627.3mV change in the drain voltage. 87

4.8 (a) The pFET voltage-adapting circuit configuration; this circuit is the autozeroing floating-gate amplifier (AFGA). (b) Response of the AFGA to an upgoing and a downgoing step input. The adaptation in response to an upward step results from electron tunneling; the adaptation in response to a downward step results from pFET hot-electron injection. This amplifier has a gain of 11.2, and I_{tun0} is 50fA. 89

4.9 (a) The pFET current-adapting circuit configuration. (b) Response of the current-adapting pFET synapse to an upgoing and a downgoing step input. This circuit configuration is unstable. 90

4.10 Simplification of small-signal models for the pFET single-transistor circuits; the nFET and source-degenerated pFET synapse circuits are simplified similarly. The top figure shows the simplified small-signal model for the drain connected to a cascode transistor. The bottom figure shows the simplified small-signal model for the drain connected to a current source. This circuit is unstable because of the negative resistance from floating-gate to ground. 91

4.11 (a) The nFET voltage-adapting circuit configuration. (b) Response of the voltage-adapting nFET synapse to an upgoing and a downgoing step input. This circuit configuration is unstable. 92

4.12 (a) The nFET current-adapting circuit configuration. (b) Response of the current-adapting nFET synapse to an upgoing and a downgoing step input. This circuit configuration is stable. 93

4.13 The behavior of the voltage autozeroing circuit using a source-degenerated pFET synapse. Unlike the nFET synapse, this circuit converges to its steady-state voltage. (a) Circuit diagram. (b) Response of this circuit to an upgoing and a downgoing step input for two different values of κ_x. The circuit behavior does not change for different channel currents if the bias drain–to–source voltage is fixed. 94

4.14 The behavior of the current autozeroing circuit using a source-degenerated pFET synapse. Unlike the pFET synapse, this circuit converges to its steady-state current. (a) Circuit diagram. (b) Response of this circuit to an upgoing and a downgoing step input for three values of drain voltage. (c) Response of this circuit to an upgoing and a downgoing step input for three different values of κ_x. 95

4.15 (a) Circuit with two pFET synapses coupled at the drain with a current source. (b) Circuit with two nFET synapses coupled at the drain with a cascode transistor. 97

4.16 Circuit diagram of the four-input source-degenerated pFET synapse. . . 98

4.17 Plot of the time-derivative of W versus W for the nFET, pFET, and source-degenerated pFET synapses. The arrows show the directions that the differential equations will take. This figure shows that the nFET and source-degenerated synapses will stabilize to the $W = 1$ steady state, while the pFET synapse will diverge from the $W = 1$ steady state. 99

4.18 The behavior of coupled pFET synapses for fixed inputs. Even though the synapse currents initially start near each other, I_1 wins and I_2 loses. I_2 decreases as a linear exponential in time due to the constant tunneling current at the floating gate. The measured I_2 saturates due to the surrounding leakage currents; the floating gate continues to increase with time. . . . 100

4.19 The behavior of coupled nFET synapses for fixed inputs. Even though the synapse currents initially start orders of magnitude apart from each other, both currents eventually converge to nearly the same steady-state level. 101

4.20 Output voltage and synapse voltage (V_1, V_2, V_3, V_4) responses due to an input step applied to the first synapse. This figure shows that the stabilizing behavior for two synapses is extendible to multiple synapses. 103

5.1 An autozeroing floating-gate amplifier (AFGA) that uses pFET hot-electron injection. The ratio of C_2 to C_1 sets the gain of this inverting amplifier. The nFET is a current source, and it sets the current through the pFET. Steady state occurs when the injection current is equal to the tunneling current. The capacitance from the floating gate to ground, C_w, represents both the parasitic and the explicitly drawn capacitances. Increasing C_w will increase the linear input range of the circuit. The capacitance connected to the output terminal, C_L, is the load capacitance. Between V_{tun} and V_{fg} is our symbol for a tunneling junction, which is a capacitor between the floating-gate and an $n$well. 106

5.2 Response of the AFGA to a 1Hz sinewave superimposed on a 19s voltage pulse. The AFGA has a closed-loop gain of 11.2, and a low-frequency cutoff at 100mHz. The 1Hz signal is amplified, but the much slower step is adapted away. 108

5.3 The small-signal model of a pFET with the effects of hot-electron injection. I assume a constant tunneling current at the floating gate (V_{fg}); this tunneling current sets the bias point for the hot-electron injection parameters. . . 109

5.4 The effect of drain-to-source voltage on the Early voltage of an nFET and pFET. The AFGA's open loop-gain as a function of the output voltage is directly related to the change in the Early voltage. The decrease in the nFET's Early voltage at high drain-to-source voltages is due to the impact ionization in its drain-to-channel depletion region. For typical steady-state output voltages around 1V to 5V, the Early voltage of both the pFET and nFET are nearly constant; therefore, the open-loop gain is constant. . . 110

5.5 Steady-state output voltage versus the tunneling voltage for three values of V_τ. This data set agrees with the model described in (5.5). The AC gain of this amplifier was 146. 111

5.6 Steady-state output voltage versus V_τ for two tunneling voltages. This data set agrees with the model described in (5.5). The AC gain of this amplifier was 146. 112

5.7 Response of the AFGA to an upgoing and a downgoing step input. The adaptation in response to an upward step results from electron tunneling; the adaptation in response to a downward step results from pFET hot-electron injection. This amplifier had a gain of 11.2. I plot the curve fits of the simplified expressions of (5.17), where either tunneling or injection dominates the restoration process. Using the curve fits, τ is 4.3s and I_{tun0} is 50fA. The value of τ can be set reliably to more than 10^5 seconds. . . 115

5.8 The response to a square wave for four different values of the tunneling voltage. This amplifier had a gain of 147; the input square wave is not shown. The steady-state output voltage decreased in the same manner as seen in Fig. 5.5 for increasing tunneling voltages. The initial tail in the upgoing response is due to the output voltage going to ground. 116

5.9 The effect of long-duration AFGA operation. (a) The responses to an upgoing and downgoing voltage step before and after 145 hours of operation. I plot the difference in the output voltage from the equilibrium DC level as a function of time; the equilibrium output voltage increased slightly over the 145 hours of operation. (b) The extracted device parameters as a function of time. Since I_{tun0}/C_2 changes more than does V_{inj}, most of the long-term change is caused from the tunneling junction, which is probably caused by oxide trapping. 118

5.10 High-frequency AFGA behavior. (a) Two AFGAs with unity gain, but with different values for C_1. The larger-capacitor circuit had $C_1 = C_2 = 300$ fF; the smaller-capacitor circuit had $C_1 = C_2 = 50$ fF. For both AFGAs, C_L was the same. I operated the two AFGAs with different subthreshold bias currents to achieve comparable settling times. (b) Two AFGAs with different gains. 119

5.11 The response of three AFGAs to the same square-wave input. All three AFGAs were identical except for C_w, and were biased by the same V_τ. Increasing C_w increases the linear range, decreases the amount of capacitive feedthrough, and decreases the low-pass cutoff frequency. 123

5.12 Measured linear range and τ_h for several unity-gain AFGAs for different C_w ratioed in units of C_1. The linear range fit is $V_{Li} = 0.063\text{V}\ C_w/C_1 + 0.125\text{V}$, and the τ fit is $\tau = 1.8\mu\text{s}\ C_w/C_1 + 2.7\mu\text{s}$. 124

5.13 An AFGA represented as a small-signal circuit. (a) The small-signal AFGA model using the small-signal pFET model. (b) The small-signal model of the effect of the noise source in the channel on the output voltage. I have neglected the effect of the gate current, as well as the Early voltage effect, in this model. (c) A simplified small-signal model of the effect of noise. For clarity, I define $R_x = \frac{C_1+C_2-C_w}{g_m C_2}$, and $C_x = C_L + C_2\left(1 - \frac{C_2}{C_1+C_2+C_w}\right)$ . . 125

5.14 Frequency response for two AFGAs with different gains. For both the high- and low-gain AFGA, $C_1 + C_2$ is approximately constant. For the high-gain AFGA, τ_l is 20mHz, and τ_h is 600Hz; for the low-gain AFGA, τ_l is 300μHz and τ_h is 40kHz. The ratio of τ_h and τ_l between the two AFGAs are equal to one-half of the ratio of the gains; the ratio is consistent with a constant $C_1 + C_2$. 126

5.15 Output noise spectrum of an AFGA with a gain of 146 for two different tunneling voltages (V_{tun}) and a constant input. The high-frequency cutoff eliminates $1/f$ noise at frequencies below 1 / $2\pi\tau_l$. The spectrum was taken for a bias current of 80nA, which corresponds to a V_τ of 0.73V. 128

5.16 Comparison of a high-gain AFGA with a unity-gain AFGA and with a generic follower-connected differential amplifier. All three amplifiers had the same V_τ voltage, and had the same bias current. The sums of C_1 and C_2 are the same for the two AFGAs. 129

5.17 Minimum and maximum output voltages versus the peak-to-peak output-voltage amplitude. The frequency of the input sine wave was 100Hz; the AFGA had a gain of 146. For small input amplitudes, the minimum and maximum output voltages symmetrically deviate from the steady-state voltage; for large input amplitudes, however, the DC output voltage follows the maximum output voltage. The DC voltage was fit to the function $0.5\ln(\mathrm{I}_0(V_{dc}$ / 1.0V)), which is equal to (5.53) with $V_{inj} = 500$mV. . . . 131

5.18 The response of an above-threshold AFGA to a downgoing step The feedback cap, C_2, of this AFGA is only the parasitic floating-gate–to–drain overlap capacitance. In the subthreshold case, the voltage linearly decreases with time; therefore, the nonlinear decrease of the output voltage for an above-threshold bias shows that the overlap capacitance changes with drain voltage. 134

5.19 Change in the AFGA output voltage with and without a continuous tunneling current. I used the same AFGA with a gain of 146 with both experiments. The experiment for no tunneling current also required that I drop the power supply; V_{dd} was set at 5V. The trace with no tunneling current was started 5 minutes after the tunneling line was dropped. 136

6.1 Circuit diagram of the autozeroing second-order section. This circuit, which is built with three autozeroing amplifiers, shows second-order behavior that is electronically controlled. 142

6.2 Step response of the autozeroing second-order section. (a) Short timescale relaxation to upgoing and downgoing input steps. The ringing of the output voltage is characteristic of a second-order system. (b) Long timescale relaxation to an upgoing input steps; a 1Hz square wave was superimposed on the input signal, and is preserved throughout the relaxation. This ringing behavior proves that the circuit exhibits at least second-order behavior from the AFGA corner frequencies set by the floating-gate currents. 143

6.3 The circuit diagram of a two-input winner-take-all circuit. 144

6.4 Time traces of the output current and voltage for small differential input current steps. (a) Time traces for small differential current steps around nearly identical bias currents of 8.6nA. (b) Time traces for small differential current steps around two different bias currents of 8.7nA and 0.88nA. In the classic WTA, the output currents would show no response to the input current steps. 145

List of Tables

4.1 Relationships of small-signal parameters 81

4.2 Values of A and B for the three synapses 97

Chapter 1 Floating-Gate Learning Systems

This thesis presents a floating-gate technology that we can use to build silicon systems that adapt and learn. **Learning** and **adaptation** have similar meanings and are often used interchangibly. Adaptation is the ability of a system to ignore baseline conditions; **adaptation** often refers to low-level processing. Examples of adaptation are gain control, habituation, and desensitization. **Learning** is the modification of system abilities in response to environmental changes; **learning** often refers to high-level processing. Examples of **learning** are aquiring motor skills, building internal representations of the external world, and so on.

Presently, biological systems are the most complex examples of both **adaptation** and **learning**; therefore, we have much to discover by examing how various engineering problems have been solved adaptively in biological systems. The biological and silicon media are similar; for example, potential barriers are established when the diffusion of charge carriers across a region of space balances the flow of carriers accelerated by the resulting electric field. The information-processing systems embodied in both silicon and biology are physical systems that are affected by the same nonidealities (mismatch, noise, etc.); biological information-processing systems perform astounding computations in spite of these nonidealities. It also seems that neurobiological systems have been optimized both for size and for power dissipation in their computational networks.

The challenge is that we cannot simply duplicate the biological solutions in the silicon media, because the constraints imposed by the biological and silicon media are not identical. It is likely that the differences in the media will lead to different optimal implementations of various information-processing systems. Mahowald's neuron circuit—an implementation of a detailed compartmental model similar to the

Hodgkin–Huxley neuron model—required 27 transistors, occupied $10^5 \mu m^2$ of circuit area, and required 11 biases [1]. We might find denser implementations, but we probably could not build such dense implementations of realistic neurons in our silicon technology.

In addition to providing us with an efficient information-processing system, building a working system based on biological principles in the silicon medium can give us a deeper understanding of the underlying phenomena. One example that my colleagues and I developed is a silicon axon [2]. This circuit is based on the biological observation that pulse widths do not diffuse away, as they do in a passive RC cable. The axon circuit uses an active mechanism to restore the pulse width; it would be interesting to make a connection to the active mechanism in the biological system.

My goal in my thesis research was to develop an analog floating-gate technology to implement adaptation and learning in silicon. Over the past 10 years, I have been developing analog integrated circuits that adapt and learn [3, 4, 5]. The approach described in this dissertation begins with the constraints that the silicon medium imposes on the learning system; letting the silicon medium constrain the design of a system results in efficient methods of computation. Analog learning systems require a large number of interacting processors; this consideration tightly constrains the size and power dissipation of each processor. This design methodology has proved essential for building analog learning systems.

1.1 Historical Perspective on Analog Implementations of Neural Systems

Analog implementations of neural systems formed a recognizable field of study roughly 10 years ago. Neural networks became an active research topic, due to the pioneering work of Hopfield [6, 7], the PDP group [8], and other researchers [9]. Since that time, use of neural networks has become an accepted numerical technique for solving many nonlinear signal-processing, optimization, and control problems, although the neu-

ral network field still has many open questions to address. Expectations for analog implementations were high, and many researchers supposed that these expectations would be fulfilled quickly. Part of the research interest in analog VLSI implementations of neural systems was fueled by digital VLSI becoming a mature field in which few solvable problems remained.

After 6 years of research, several noteworthy neural systems emerged. Biologically inspired neural systems were built in three major categories, mostly as a result of the work of Mead's group and of other related laboratories. Image-processing systems, which were the first active-pixel chips, include silicon retinas [10, 11, 12, 13] optical-flow processors [11], and stereo processors [14]. Auditory-processing chips included many cochlea chips [15, 16], an auditory-localization chip [17], and a chip for binaural hearing [18]. These systems significantly advanced the state-of-the-art, but the majority did not include the additional complexity of adaptation or learning.

Connectionist neural systems are typically built as direct mappings of mathematical and computer models of neural networks into analog silicon hardware. Intel's ETANN chip was the first commercially available neural-network integrated circuit, and the ETANN used floating gates for weight storage [19]. The implementation of the Heuralt–Juetten algorithm by Andreou's group was one of the most successful large-scale adaptive neural systems, but it required a great deal of circuit complexity [20]. Other researchers implemented unsupervised learning and back-propagation algorithms with mixed success [21, 5, 22]. The successful analog implementations of connectionist networks included algorithmic modifications that facilitate its implementation in silicon [23, 5, 24]; history has shown that the success of an implementation is strongly correlated to the degree to which the algorithm is adapted to the silicon medium.

A substantial portion of the early work in neural-network implementations went into developing dense multiplier circuits. Currents are preferred for outputs, because the summation typically required for most connectionist models is easily performed on a single wire, and voltages are preferred for inputs because they are easy to broadcast. Since an input voltage should modulate an output current, most implementations

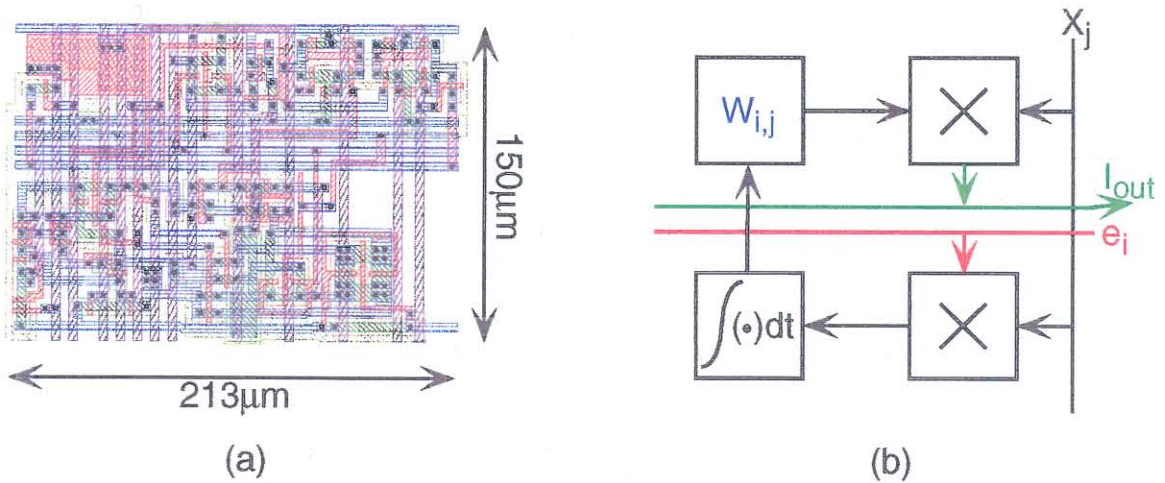


Figure 1.1: In previous neural network implementations, the adaptive elements (synapses) were complex and required a tremendous amount of area. (a) Synapse layout from [5]; the size of this synapse was typical of dense synapse designs of similar complexity. (b) Typical implementations used separate computation, memory, and adaptation blocks, because these networks were direct implementations of mathematical models.

employ a variable resistance or transconductance element. The approaches include synapses based on

- Fixed resistances, which were the earliest implementations [25],
- Switched-capacitor, charge-coupled devices (CCD), and related implementations [26, 3],
- Gilbert multiplier cells [27],
- Linearized conductance elements [28, 29, 30, 31].

Murray [32] and other researchers argued that pulse computation—which can include mean firing rate computation, pulse-width modulation, and related techniques—would make multipliers denser than analog-valued multipliers; they were wrong. Pulse computation may prove important in biologically inspired implementations that employ event-based processing, since biological systems communicate and compute with action potentials that encode events.

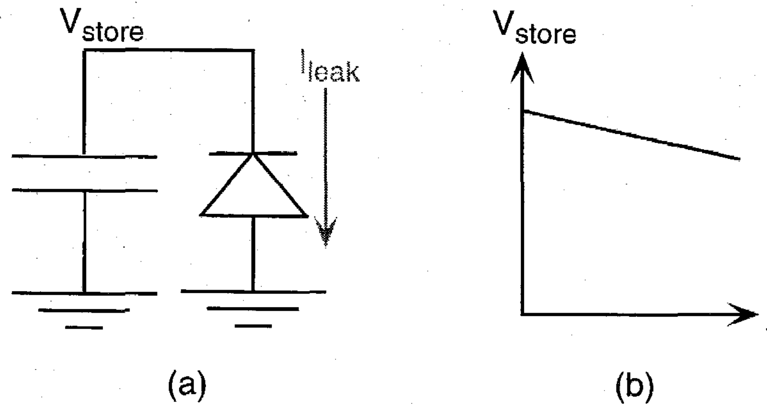


Figure 1.2: Transistor leakage currents limit adaptation time-constants. (a) Circuit schematic when storing a voltage on a capacitor. (b) Plot of the resulting stored voltage versus time. The constant leakage current decreases linearly the stored voltage over time.

Even though many researchers made progress over this 6-year period, the resulting analog neural-network implementations still had many shortcomings. Most networks were direct implementations of mathematical models, and therefore the synapses required large circuit complexity, as illustrated in Fig. 1.1. Not only does large circuit complexity consume tremendous circuit area and power, but also the chance of a network operating correctly decreases exponentially with cell size. One reason that the synapses were large is that the implementations used separate memory, computation, and adaptation blocks. In addition, few successful systems had been built that included adaptation, and even fewer such networks had been built elegantly.

The effect of p-n junction leakage currents, illustrated in Fig. 1.2, was the most difficult problem to overcome to build efficient adaptive circuits. Real-time adaptation and learning requires time constants in the range from 10ms to days; because of junction leakage currents, most integrated-circuit processes are restricted to time constants shorter than 1s unless they use prohibitively large capacitor areas. The adaptation rate must be much smaller than the input-signal rate; therefore, the addition of adaptation to a system constrains the minimum computation rate. A system that has several levels of adaptation requires that the slowest time constant be several orders of magnitude slower than the slowest input-signal rate. If the input-signal fre-

quencies are relatively fast—measured in MHz frequencies—then adaptation can be employed directly in these systems. The adaptive photoreceptor circuit is one of the best examples of adaptive non–floating-gate circuits; this circuit amplifies only fast changes in the input (ms timescales), but does not amplify the background intensities (100ms timescales) [33, 34]. The *Tobi element*, used in the hysteretic differentiator, is the key component that performs the adaptation. Even this circuit shows the time-constant restrictions, and higher-order adaptation is not available.

Four years ago, Brad Minch, Chris Diorio, and I, with Carver Mead, set out to develop a natural silicon technology for learning and adaptation based on the constraints of the silicon medium. We collaborated closely to actualize the possibilities of this new technology. This dissertation is a detailed account of the work that I contributed to this collaboration. In addition, this collaboration produced several related publications [35, 36, 37, 39, 40, 41, 42, 43], as well as the dissertations of my two colleagues. Our collective work provides a good example of synergistic cooperation; none of us would have accomplished nearly as much without the assistance of the others.

1.2 Floating-Gate Technology

To use our medium efficiently, we must find its computational primitives. Digital MOS fabrication processes have been designed and optimized to produce high-quality MOSFET (Metal-Oxide-Semiconductor Field-Effect Transistor) devices with well-controlled gate oxides. The CMOS (Complementary Metal Oxide Semiconductor) process also yields several devices that CMOS process engineers did not intended explicitly; this includes the floating-gate MOSFET. Although floating-gate devices that are fabricated in highly specialized processes have long been used for nonvolatile memories, other potential applications of floating-gate devices have remained largely unexplored. Inspired by the biological systems, we have utilized physical characteristics of the silicon medium that have traditionally posed problems for engineers.

The ETANN chip— one of the few neural networks built with floating-gate technology—

used electron tunneling to program the stored weights from off chip (it had no online adaptation) [19]. Concerns about this implementation fueled several arguments against floating-gate implementations. First, most researchers assumed that specialized EEPROM (Electrically Erasable Programmable Read Only Memory) processes, which are not available to most researchers, were necessary for programming the synapses. This concern was partially dispelled by an implementation of interpoly-oxide electron tunneling in the 2μm Orbit process [44]; poor quality of interpoly oxide resulted in poor matching and limited repeatability of tunneling-current behavior. We developed our floating-gate technology from the same standard CMOS processes. The voltages required for electron tunneling (and hot-electron injection) exceed the rated supply voltage, and therefore would be dangerous engineering. Of course, CMOS process designers have considered electron tunneling (and hot-electron injection) to be the source of many digital VLSI reliability problems, and have not seen it as a potentially useful phenomenon. Floating-gate memories were difficult to program with high precision due to complicated programming schemes and nonlinear tunneling characteristics. Floating-gate memories typically require long programming times (when compared with digital clock rates); this slow behavior has been considered by many researchers as a serious limitation of this technology for neural systems. These floating-gate memories, like EEPROM devices, required that the floating gate be driven over a large voltage range to tunnel electrons onto the floating gate; synaptic computation must stop for this type of weight update.

Consider the potential of this floating-gate technology from an adaptive-systems perspective. The reason this technology is well suited for implementing adaptation is that we are not building analog EEPROMS, but rather circuit elements with important time-domain dynamics. In other words, floating-gate devices are not used just for memories anymore. Figure 1.3(a) shows a cross section of a floating gate, which is a polysilicon gate surrounded by SiO_2. Charge on the floating gate is almost permanently stored, because it is completely surrounded by a high-quality insulator. Consequently, once the adaptation is finished, the resulting network state is preserved nearly indefinitely. The floating gate can modulate a channel between a source and

drain, as shown in Fig. 1.3(b). Since the floating-gate voltage can modulate a MOSFET's channel current, the floating gate not only serves as a memory, but also can be an integral part of a computation. In subthreshold, the channel current is proportional to the exponential of the floating-gate voltage; therefore, these computations are a direct function of the stored charge on the floating gate. We showed that we can use these floating-gate devices to compute generalized translinear functions by a particular choice of capacitive couplings between the floating gates [38]. The most well-known translinear circuits are multipliers, which date back to the Gilbert multiplier [47].

Fig. 1.3(c) shows electrons moving from the transistor channel to the floating gate by hot-electron injection. The floating-gate charge can be modified by electron tunneling, by hot-electron injection, and by ultra-violet photoinjection. The oxide currents can be arbitrarily smaller than the MOSFET's channel current; typically, the oxide currents are at least 10^4 times smaller than the channel currents. These small current levels gives us *slow* programming rates, which are precisely what we need to build adaptive systems, because the adaptation should integrate over many presentations of the inputs, or over several periods of the input. Continuous-time adaptation occurs because the gate current is a function of the device parameters; therefore, floating-gate adaptation is a direct function of the computations being performed. Hot-electron injection is an important adaptation mechanism, because floating-gate current is proportional to channel current and to an exponential function of drain voltage. In subthreshold MOS transistors, the adaptation time constants due to the gate currents well match the computation time constants due to the transistor channel currents.

A frequently asked question is "What is the resolution of a floating-gate device?" The question implies that we are still talking about a memory element. At a fundamental level, the precision is limited by the number of voltage levels on the floating gate, where the voltage level is determined by the voltage that a single electron generates on the floating-gate capacitor. The limiting factor, as it is in non–floating-gate neural memories [45], is the circuitry that converts the stored representation to the

required circuit value. We have shown that it is the analog circuitry, rather than the floating gate, that limits analog resolution in our processes [41].

1.3 Thesis Overview: Electron Transport to Floating-Gate Circuits

This dissertation systematically presents a bottom-up progression from electron transport to floating-gate circuits. Unless otherwise stated, the data presented are experimental results that I measured from integrated circuits fabricated in either the 1.2μm or 2μm, double-poly $n$well Orbit CMOS process available through MOSIS. We will begin in Chapter 2 with my model of hot-electron injection and impact ionization derived from a first principles model of hot-electron transport. This analytical model of hot-electron transport in the depletion region between the channel and drain regions was derived from Boltzman transport. From a physics perspective, this model is important because it explains the underlying phenomena of both impact ionization and hot-electron injection, and complements the current approach of semiconductor modeling of Monte Carlo numerical calculations using full-bandstructure models. From a circuits perspective, we have the first realistic device model of hot-electron injection that we can use to develop, simulate, and verify floating-gate circuits.

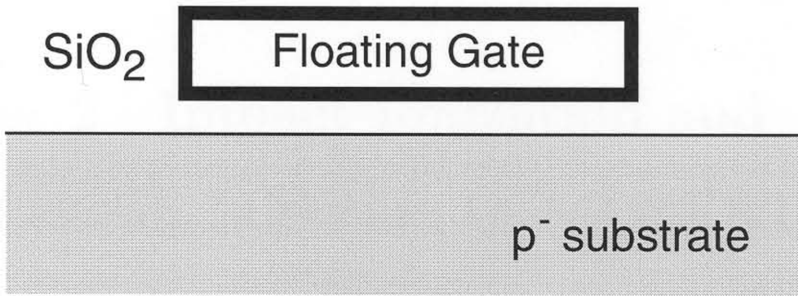
Chapter 3 presents the single-transistor learning synapses, which are the fundamental circuit elements for adaptation and learning circuits. The single-transistor synapse, invented and developed by me and my collaborators, Diorio and Minch, comprises a hot-electron–injecting floating-gate transistor and a well-tunneling junction. This chapter describes the basic nFET and pFET synapses; other variations that we have invented will be presented in Diorio's dissertation. These synapses simultaneously store the analog weight, compute a product based on this weight and on the input signal, and adapt this weight, all in the same element. This chapter briefly discusses electron tunneling, a well-understood phenomenon that we use to remove electrons from the floating gate. This chapter also presents a simplier device-level

model of hot-electron injection in nFET and pFETs. The single-transistor-learning synapse represents the transition from devices to circuits for this floating-gate technology. The small size and low-power operation of single-transistor synapses permits us to develop dense synaptic arrays. When the steady-state source current is used as the representation of the weight value, both the incrementing and the decrementing functions are proportional to a power of the source current.

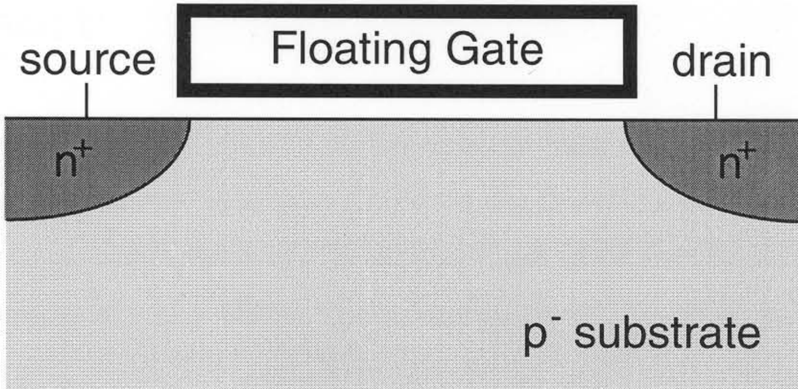
Chapter 4 presents the possible dynamics of simple continuous-time floating-gate circuits. Circuits that have one nFET or pFET synapse show a range of stabilizing and destabilizing behaviors, because each circuit imposes a particular type of feedback to the floating gate. By providing feedback to the source, we get another synapse type with unique dynamics, because voltage changes in both the floating gate and drain stabilize the floating gate. Multiple floating-gate circuits show both competitive and cooperative behaviors between synapses, consistent with Hebbian and anti-Hebbian learning. Modeling the dynamics of a floating-gate circuit gives us intuition into developing more complex circuits, and shows the strengths of simple floating-gate current models.

Although the single-transistor synapse can serve as the fundamental adaptive element, and although its feedback properties are well characterized, we still need to demonstrate that an adaptive circuit based on our floating-gate technology could learn continuously. Chapter 5 presents the autozeroing floating-gate amplifier (AFGA), a floating-gate amplifier that adapts its output voltage back to the same steady-state value over a long timescale. The AFGA is the simplest practical form of a floating-gate circuit that adapts. This continuous-time bandpass amplifier uses capacitive feedback and capacitor ratios to set several important circuit properties. The AFGA circuit is the basic building block for a family of continuous-time amplifier and filtering circuits; it provides an attractive low-power alternative to switch-capacitor filters.

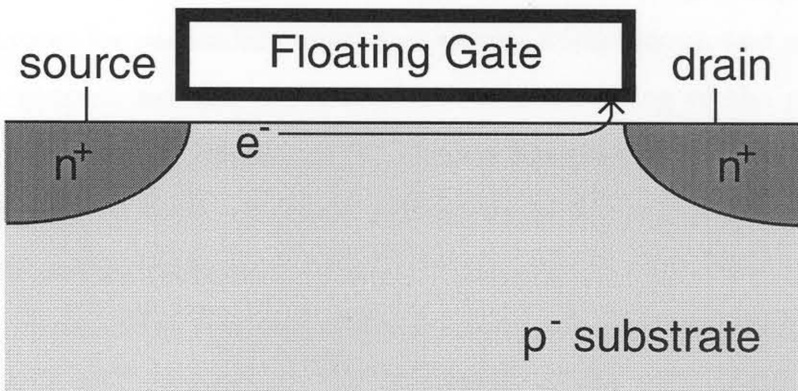
Chapter 2 starts by considering my first-principles model of hot-electron injection. The circuit-oriented reader might choose to skim read Chapter 2, and then jump to Chapter 3.



(a)



(b)



(c)

Figure 1.3: Illustration of our floating-gate CMOS technology. (a) Charge on the floating-gate is nearly permanently stored because electrons are surrounded by a high quality insulator. (b) Since the floating-gate voltage can modulate a MOSFET's channel current, the floating gate is not only a memory device, but also can be an integral part of a computation. (c) The floating-gate charge can be modified by electron tunneling and hot-electron injection. Continuous-time adaptation is possible if the floating-gate current is a function of other device parameters. In this respect, hot-electron injection is an important adaptation mechanism, because floating-gate current is proportional to channel current and an exponential function of drain voltage.

Chapter 2 Impact Ionization and Hot-Electron Injection in MOSFETs Derived Consistently from Boltzman Transport

Floating-gate MOS devices [48] rely on hot-electron injection and tunneling phenomena to control the amount of charge stored on the floating gate. In this traditional mode of operation, employed in EEPROM and Flash ROM memories [49, 50], the device is used simply as a nonvolatile, charge-storage node. Another application of this device is in analog silicon synapses, where the device stores analog nonvolatile charge while performing a single-quadrant multiplication on the applied voltages and adapting the floating-gate charge as a function of the applied voltages [35, 36]. Optimization of silicon synapses for particular circuit and system applications, and of floating gate structures in general, necessitates a thorough understanding of the physical mechanisms involved in device operation. To explain hot-electron injection in SiO_2, one must understand how electrons obtain sufficiently high energies that they overcome the silicon–silicon-dioxide barrier.

Shockley introduced the concept of a *lucky-electron* in modeling the impact-ionization process in silicon and germanium [51]. According to the lucky-electron model, the probability (P_{nc}) that an electron does not experience collisions, and thus does not lose energy, is proportional to an exponential function of the ratio of distance that the electron travels (x) and its mean free path (λ):

$$P_{nc} \propto e^{-x/\lambda}. \tag{2.1}$$

Scientists employed this lucky-electron model to explain hot-electron injection in MOS

capacitors [52]: Shockley's simple theory was in good agreement with the data over a wide range of dopings and temperatures. The same model was also employed to explain hot-electron injection and impact ionization currents in MOS transistors biased above threshold [53, 54]. Tam and colleagues observed a correlation between the hot-electron injection currents and the impact-ionization currents in MOS transistors [53], but provided no physical connection between these two phenomena. Although the lucky-electron theory provides simple explanations, it does not give insights on the actual physical processes, and it cannot be related to an underlying hot-electron distribution function.

A second modeling approach assumes an a priori hot-electron distribution function [52, 54]. Usually, the distribution function (f) as a function of energy (E) and an effective temperature (T_c) is

$$f(E) = e^{-E/kT_c}. \tag{2.2}$$

However, the results of Monte Carlo simulations [55, 56] suggest that, at the energies at which impact ionization and hot-electron injection occur, a constant electron temperature is not a good approximation. Baraff [57, 58] proposed an analytic model for impact ionization deriving from a spatially-uniform, Boltzmann transport equation. His model fits impact-ionization data from p-n junctions, and in some regimes, agrees with the lucky-electron theory.

Unfortunately, we need a model derived from a spatially-varying Boltzmann transport equation, because the distribution function in the drain-to-channel depletion region in a MOS transistor is far from spatially uniform for significant impact-ionization and hot-electron-injection currents. When the impact-ionization and hot-electron-injection currents in a MOS transistor are significant, the distribution function in the drain-to-channel depletion region is far from spatially uniform; therefore necessitating a model derived from a spatially-varying Boltzmann transport equation. As a result of the added complexity in the Boltzmann transport equation little progress has been made towards analytically relating the description of the electron distribution functions to the measured data of hot-electron-injection currents and impact-ionization

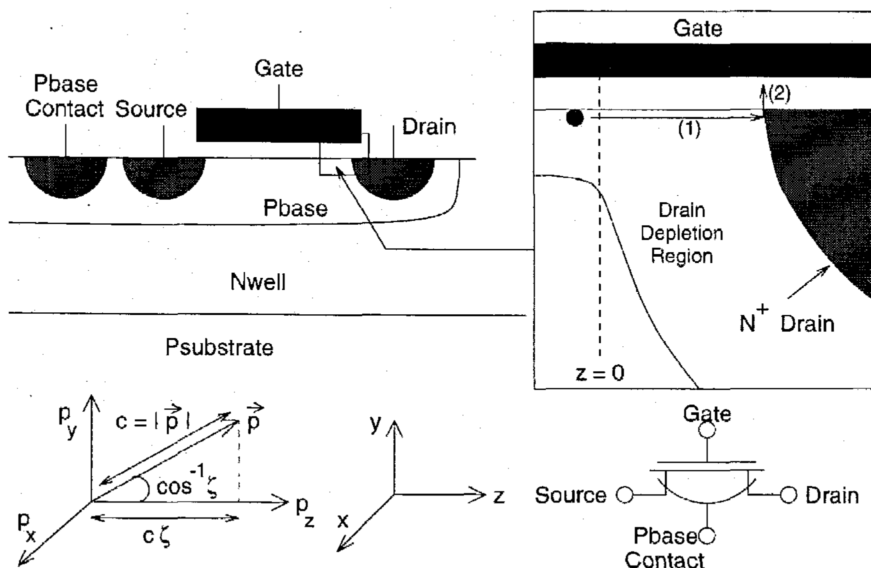


Figure 2.1: Cross section of the MOSFET device we used to measure the hot-electron effects. It uses a moderately doped ($1\text{x}10^{17}\text{cm}^{-3}$) substrate to achieve a high threshold voltage which allows hot-electron injection for bias current levels in subthreshold. The higher doping is consistent with a 0.2μm channel length CMOS process; therefore the effects are directly applicable to modern processes although the device was fabricated in a 2μm process. The *n*well isolates the moderately doped substrate region from the surrounding substrate, and allows measurement of substrate current. Holes resulting from impact ionization are measured at the *p*base contact. The hot-electron injection process is identical for the FET with or without the isolating *n*well. Inset: the electron is accelerated through the drain depletion region (1), and when it gains energy greater than the Si–SiO_2 barrier, the electron is injected over the Si–SiO_2 barrier to the floating-gate (2). Lower Left: Circuit symbol for the *n*FET with the *p*base implant.

currents. For a level of description beginning with electron distribution functions, the only models that agree with experimental data of both hot-electron injection and electron tunneling are numerical Monte Carlo models [56].

In this chapter, I develop a quantitative model of the impact ionization and hot-electron processes, that is derived consistently from a single spatially varying hot-electron distribution function. The hot-electron transport in the drain-to-channel depletion region is modeled using the spatially varying Boltzmann transport equation. I also show an analytical relationship between the impact ionization in the drain-to-channel depletion region of an MOS transistor and hot-electron injection on the gate.

I verified the model experimentally using the MOS transistor structure with a moderately high substrate doping (1×10^{17} cm^{-3}) shown in Fig. 2.1. I simultaneously measured substrate and gate currents from the device biased in subthreshold. Under subthreshold biasing conditions, the surface potential is fixed; therefore, the high field effects are confined to the drain-to-channel depletion region where carrier concentrations are at low concentrations. I shall show analytically that, at the energies where impact ionization and hot-electron injection occur, a constant electron temperature is not a valid assumption. The analytic model of impact-ionization and hot-electron injection currents agrees well with experimental data.

This chapter is structured as follows. Section 2.1 formulates a Boltzmann Transport Equation and makes reasonable device approximations to model the transport in the drain-to-channel depletion region. Section 2.2 solves analytically for the distribution function in a self-consistent manner. I solve for the average trajectory taken by a hot electron as a function of position through the depletion region. This determines the average electron energy and direction as a function of position. In this coordinate system, phonon collisions act to diffuse the distribution function as a function of position, and impact ionization collisions remove electrons from participating in hot-electron injection. A new collision operator for impact ionization is proposed, which is necessary to simultaneously fit the hot-electron injection and impact ionization data. Section 2.3 uses this distribution function to calculate the probabilities of impact ionization and hot-electron injection as a function of channel current, drain voltage, and floating gate voltage. I compare my analytical model to measurements of hot-electron injection and electron impact ionization; the model simultaneously fits both sets of experimental data. A discussion follows in Section 2.4.

2.1 Electron Transport in the Drain to Channel Depletion Region

The Boltzmann Transport Equation determines the spati-temporal evolution of an electron distribution function under low to moderately high fields. Even though it can be argued that its use in this work is questionable because of the high electric fields involved, I believe that it is a good starting point. Monte Carlo simulations of impact ionization in silicon [56] indicate that the collision times in the high energy regime of concern here, are at most a factor of two larger than 10fs. In other words, the time between collisions is on the same order of magnitude as the time of a collision. The Boltzmann transport equation is still approximately valid when the average time between collisions is on the order of 10fs; therefore, its use is justifiable. The alternate and more rigorous approach is to model this problem using quantum transport techniques [59]; this will be the natural followup to this work. In addition, Section 2.4 will discuss the effects of zeroth-order quantum effects, like collision broadening.

2.1.1 Boltzmann Transport Under an Applied Electric Field

The density of an electron gas in 6-dimensional space is defined by the distribution function, $f(x, y, z, p_x, p_y, p_z)$, where p_x, p_y, and p_z are the components of momentum in x, y, and z coordinates. Integrating $f()$ over all six coordinates equals a constant, which without loss of generality is is set equal to one. The evolution of this distribution function under the influence of applied fields is described by the Boltzmann transport equation [60]

$$\frac{\partial f}{\partial t} + \vec{v} \cdot \nabla f + q\vec{\mathcal{E}} \cdot \nabla_{\vec{p}} f = S(f), \tag{2.3}$$

where $\vec{v}$ is the average velocity vector, $\vec{\mathcal{E}}$ is the applied electric field, and the function $S(f)$ is defined as the difference of the rates of electrons being scattered into and out of this six dimensional space. The collision operator, $S(f)$, if formulated as [60]

$$S(f) = \left(\frac{\partial f}{\partial t}\right)_{scatterin} - \left(\frac{\partial f}{\partial t}\right)_{scatterout}. \tag{2.4}$$

2.1.2 Simplifications for Hot-Electron Transport in the MOSFET Drain-to-Channel Region

We can simplify the general Boltzmann transport formulation by making the following simplified approximations to the carrier transport in a subthreshold MOSFET. This work is concerned with hot electron injection in applications where the time scales involved are in the μs range or slower. Since scattering times are on the order of ps and fs, we can safely assume that the electron distribution function has sufficient time to reach its steady state. The electron gas is due to several electrons in time, although at any given point in time there will be very few electrons in the drain-to-channel depletion region.

I will assume that the 2D confinement in the MOS inversion layer only slightly perturbs the density of states, and to a first order does not affect our solutions. The electric field from the charge on the gate confines the electrons to a region near the substrate surface. A triangle barrier of height 2kT and a width of 10nm is just on the edge of the regime where the continuum approximation for the energy levels is still valid. Monte Carlo calculations in [61] indicate this may not always be the case.

Under normal operating conditions there is no electric field in the x or width dimension of the transistor, and therefore the electric field vector has a component only in the z direction, which I write as

$$\vec{\mathcal{E}} = (\mathcal{E}_x, \mathcal{E}_y, \mathcal{E}_z) = (0, 0, \mathcal{E}) \tag{2.5}$$

The electrostatic potentials in the y-direction result in small electron oscillations near the Si-SiO_2 interface; therefore, the electron's y-velocity distribution is uniform over a given region in x and z, and the electron's maximum y-displacement is negligible. Since collisions restore electron distributions to equilibrium, the electron distribution in the y direction will remain uniform. I will further discuss the electron confinement and the implications of the y-direction electron distribution in Section 2.4. Under the

above simplifying assumptions, (2.3) simplifies to

$$\left(v_z\frac{\partial}{\partial z}+q\mathcal{E}\frac{\partial}{\partial p_z}\right)f=S(f) \tag{2.6}$$

where ∂p_y and ∂y terms dropped out due to the confinement of the electrons in the y direction, and ∂x term dropped out due to the uniform distribution function along the width of the device.

Following Baraff's approach [58], I choose a coordinate system that orthogonalizes the total momentum and the average direction of that momentum. By transforming the momentum variables into polar coordinates (see Fig. 2.1), we get

$$\zeta\frac{c}{m^*(c)}\frac{\partial f}{\partial z}+\zeta q\mathcal{E}\frac{\partial f}{\partial c}+q\mathcal{E}\frac{1-\zeta^2}{c}\frac{\partial f}{\partial \zeta}=S(f) \tag{2.7}$$

where c is the magnitude of the average momentum vector, and ζ is the cosine of the angle of $\vec{p}$ and the z axis. ζ describes the degree of anisotropy of the distribution function. I define E, the electron energy, as

$$E(c)=\frac{c^2}{2m^*(c)} \tag{2.8}$$

where $m^*(c)$ is the effective mass of the electron. Substituting (2.8) into (2.7) we can get the following expression for the simplified BTE in this energy coordinate system as

$$\frac{\partial f}{\partial z}+q\mathcal{E}\frac{\partial f}{\partial E}+q\mathcal{E}\frac{1-\zeta^2}{\zeta E}\frac{\partial f}{\partial \zeta}=\frac{m^*(c)}{\zeta c}S(f). \tag{2.9}$$

The collision operators discussed in the next section are given in terms of $\frac{m^*(c)}{c}S(f)$, which effectively allows the formulation of a model that does not address explicitly the details of band structure. In the two subsections, we will define the collision operators in terms of $\frac{m^*(c)}{c}S(f)$, which will allow us to solve for the distribution function without knowing the detailed silicon bandstructure.[1]

[1] Canceling out the effects of the bandstructure may limit the predictive power of this model. This insight by Karl Hess is appreciated.

2.1.3 Collision Operators

The collision operator, (2.4), was introduced to describe electron scattering. The main scattering mechanisms in silicon are:

- **Phonons** are quanta of lattice vibrations [60], and are divided in two groups. *Acoustical phonons* are low energy phonons whose energy is proportional to the magnitude of their momentum vectors. *Optical phonons* have energies much larger than acoustical phonons, and these phonons have nearly constant energy as a function of its momentum. This energy is designated by E_R.

- **Ion-impurity** scattering is an elastic scattering mechanism, where a collision does not result in a change in energy of the electron, but only a change in momentum direction. At high energies, the deflections are either small in angle or lead to impact ionization, therefore we will neglect this mechanism except to explicitly model the impact ionization collisions.

- **Impact ionization** collisions occur when a high-energy carrier (electron or hole) can give up sufficient energy to liberate another electron into the conduction band, which also creates a hole in the valence band. This mechanism has been considered in several references [51, 56, 57, 62, 63].

- **Electron–electron** scattering occurs only in devices with a high density of electrons; therefore, this mechanism is negligible in the depletion regions and for subthreshold current levels because of the low density of electrons [64] .

Since only optical phonon collisions and impact ionization processes will sufficient affect the electron's energy, my model of the collision process will include these two mechanisms.

We shall proceed by first deriving a model for optical phonon collisions. Figures 2.2 and 2.3 graphically shows the effect of a phonon collision. I model optical phonon

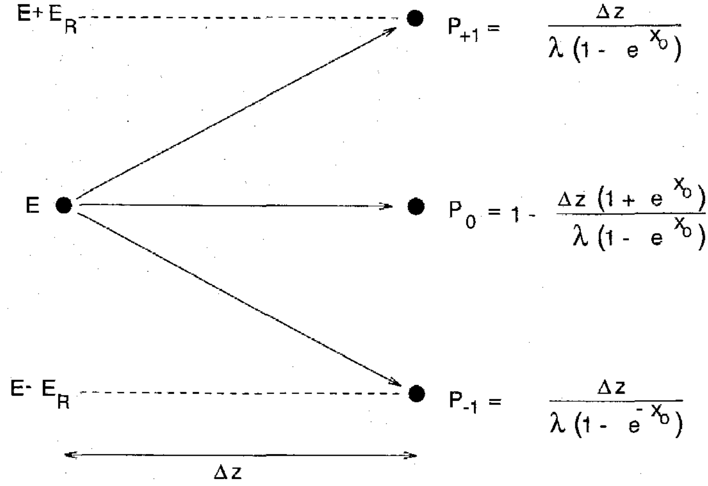


Figure 2.2: Graphical representations of the phonon collision operator. An electron can either gain or lose energy due to a phonon. The transition probabilities of gaining or losing energy due to a phonon are different. The operator is based on a mean free path of a collision; therefore the longer an electron goes in a region, the more likely it will gain or lose energy due to a phonon collision. Gaining or losing a phonon will alter the electron's momentum as well as its energy.

scattering as

$$S(f)_{op}\left[\lambda(e^{\frac{E_R}{kT}}-1)\frac{m^*(E)}{c(E)}\right] = \begin{array}{l}\left(1+\frac{E_R}{E}\right)^{1/2}\left(e^{\frac{E_R}{kT}}f(E+E_R)-f(E)\right)+\\ \left(1-\frac{E_R}{E}\right)^{1/2}\left(f(E-E_R)-e^{\frac{E_R}{kT}}f(E)\right),\end{array} \tag{2.10}$$

which was derived elsewhere [60]. Assuming no other forces on the electron, the phonons change the distribution function after the electron travels a small distance (Δz) away. In Fig. 2.2, I graphically show the equation:

$$f(z,E) = f(z+\Delta z,E)\left(1-\frac{\Delta z}{\lambda}\frac{1+e^{x_o}}{1-e^{x_o}}\right)+\frac{\Delta z}{\lambda}\frac{e^{x_o}}{1-e^{x_o}}f(z+\Delta z,E-E_R)$$

$$+\frac{\Delta z}{\lambda}\frac{1}{1-e^{x_o}}f(z+\Delta z,E+E_R), \tag{2.11}$$

where $x_0 = \frac{E_R}{kT}$; the different terms for absorbing and emitting a phonon arise from the phonon density of states. The interaction with a phonon increases or decreases the electron energy by E_R, which I set to the known silicon value of 63meV [51]. On

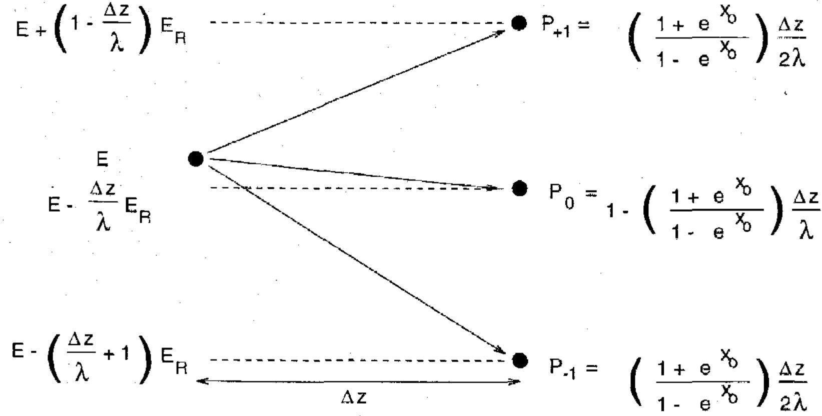


Figure 2.3: A second way to graphically represent the phonon collision operator is considering an electron either gaining or losing energy due to phonons with equal strength along a particular coordinate system.

the average, an electron makes a collision with a phonon in a distance λ, or the mean free path of the electron; I use $\lambda = 6.5$nm because it agrees with my experimental data and with previous results [51, 57].

To gain more insight on the physical mechanisms involved, I simplify (2.10) by performing a Taylor series expansion for $E >> E_R$

$$S_{op}(f)\left[\frac{m^*(E)}{c(E)}\right] \approx F(T)\frac{E_R^2}{2\lambda}\frac{\partial^2 f}{\partial E^2} + \frac{E_R}{\lambda}\frac{\partial f}{\partial E}, \tag{2.12}$$

where $F(T)$ is

$$F(T) = \frac{e^{\frac{E_R}{kT}} + 1}{e^{\frac{E_R}{kT}} - 1}, \tag{2.13}$$

which is nearly equal to one at room temperature ($T = 300K$). This expansion is valid to fourth order in derivatives in energy, and is very nearly equal to the original expression for energies greater than a few E_R. In Fig. 2.3, I show this transformation graphically, by modeling the asymmetric phonon scattering event as an average energy loss (corresponding to $\frac{\partial f}{\partial E}$) and a symmetric scattering event around this resulting energy (corresponding to $\frac{\partial^2 f}{\partial E^2}$). A similar expansion and simplification has been done for polar optical phonons [65].

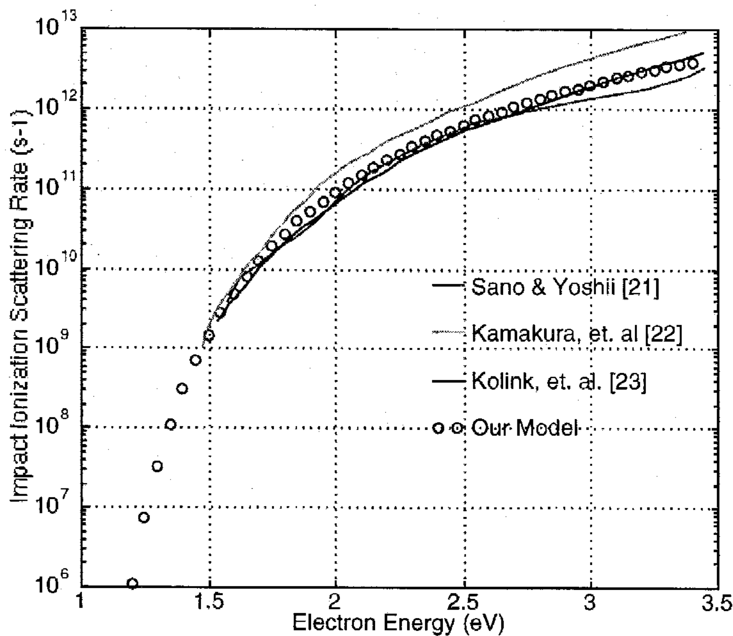


Figure 2.4: Plot of previous calculations of impact-ionization rate versus electron energy in silicon, and our derived impact-ionization rate from our measured impact-ionization and hot-electron injection data. I have assumed a constant velocity, since our model measures the impact-ionization mean-free length. Our measured data is directly related to $L(E)$ and not τ_{ion}.

Next, I will derive the collision operator for impact ionization. The collision operator for impact ionization has the general form of

$$S(f)_{ion} = -\frac{f}{\tau_{ion}(E)} = -\frac{c(E)}{m^*(E)}\frac{1}{L(E)}f, \tag{2.14}$$

where τ_{ion} is the mean free time for an impact ionization collision, and L(E) is the mean free path, which is a function of the electron energy. The first impact-ionization models were of the form [51, 57, 62]

$$\frac{1}{\tau_{ion}} \propto H(E - E_{th}), \tag{2.15}$$

where $H()$ is the unit step function or the "Heavyside Operator", and E_{th} is the threshold energy for impact ionization. Keldish [63] proposed the impact-ionization model:

$$\frac{1}{\tau_{ion}} \propto (E - E_{th})^2 H(E - E_{th}). \tag{2.16}$$

Wolfe [62] used E_{th} equal to 2.3eV; other researchers [51, 57, 58] used E_{th} equal to the bandgap of silicon ($E_{th} = 1.1$eV) since the bandgap is the minimum amount of energy required for an impact-ionization event. Monte Carlo simulations [56] of impact ionization and hot-electron injection using (2.16) showed that a small ionization rate and $E_{th} = 1.1$eV best fit the data; the resulting ionization rate was nearly constant only for energies greater than 2.3eV.

Recently, several researchers put considerable effort in calculating the impact-ionization collision times directly from band-structure calculations, and then using these results with some success in Monte Carlo simulations of hot-electron phenomena [66, 67, 68]. Figure 2.4 shows several numerically computed impact-ionization rates. Based on these calculations and our experimental measurements of impact ionization and hot-electron injection, we propose the following energy dependence for the impact-ionization mean free length:

$$L(E) = l_{ion} \exp\left(\sqrt{\frac{119rmeV}{E - .95\text{eV}}}\right). \tag{2.17}$$

Figure 2.4 shows our functional form with these three numerically calculated models, where l_{ion} is 1.81×10^{-2} nm. I have assumed a constant velocity of 8.1×10^6 cm/s in converting from $L(E)$ to τ_{ion}, since our measured data is directly related to $L(E)$ and not τ_{ion}. This functional form is a curve fit to experimental data of $L(E)$ derived from my experimental measurements of hot-electron-injection and impact-ionization currents in Section 2.3.

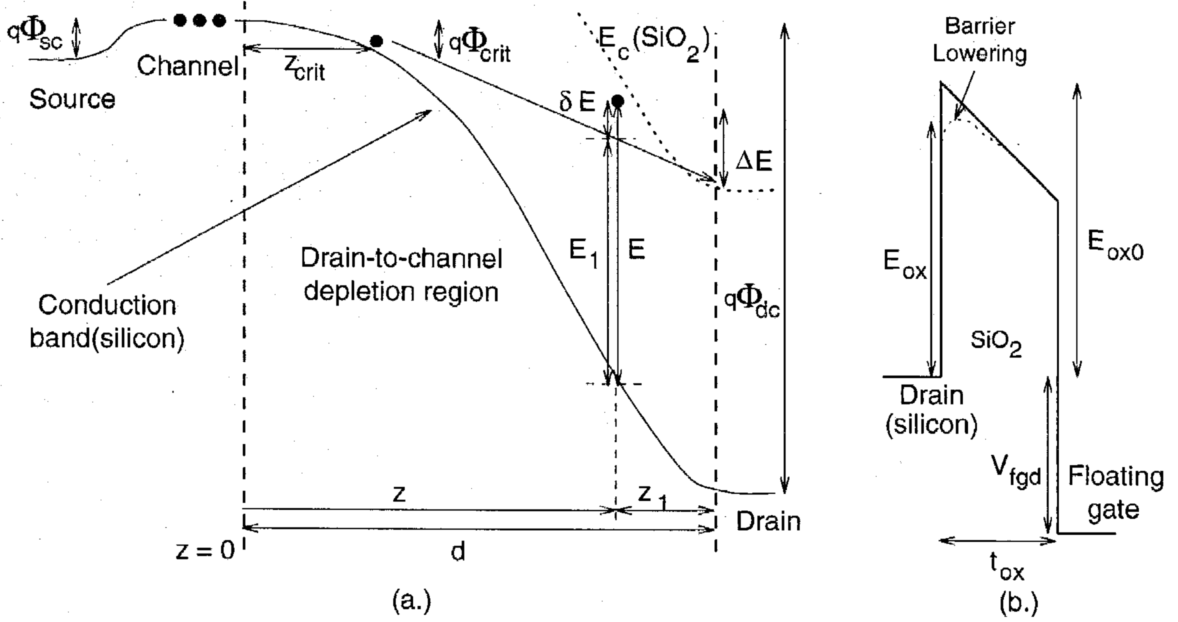


Figure 2.5: Band diagram illustrating hot-electron injection in a MOSFET biased in subthreshold. The appropriate variables in the Boltzmann transport equation and its variable transformations are shown on the graphs. (a) Band diagram along the surface of the Si-SiO_2 barrier. This region is the lowest local potential in either material; therefore the electrons are most likely to travel along this path. This region corresponds to path (1) in the inset in Fig. 2.1. (b) Band diagram at the drain edge from the substrate to the gate. This region corresponds to path (2) in the inset in Fig. 2.1.

2.2 Solving the Electron Distribution Function

This section solves for $f(z, E, \zeta)$ from the Boltzmann transport equation. Under the simplifying assumptions of the previous section, the resulting Boltzmann transport equation for $f(z, E, \zeta)$ can be written as

$$\frac{\partial f}{\partial z} + \left(q\mathcal{E} - \frac{E_R}{\lambda\zeta}\right)\frac{\partial f}{\partial E} + q\mathcal{E}\frac{1-\zeta^2}{\zeta E}\frac{\partial f}{\partial \zeta} = F(T)\frac{E_R^2}{2\lambda\zeta}\frac{\partial^2 f}{\partial E^2} - \frac{f}{L(E)} \tag{2.18}$$

with the parameters defined in the previous section. I begin by seeking a solution for the energy and momentum angle of an average electron as a function of position. I define the energy of an average electron as $E = E_1(z)$ and we define the average momentum angle of the average electron as $\zeta = \zeta_1(z)$. I then transform (2.18) around this average electron energy path. When impact ionization does not occur, we will

see that the distribution function satisfies a diffusion equation along this path. When impact ionization occurs, we will see that the solution is a product of the solution to a diffusion equation and a loss term due to impact ionization.

Figure 2.5 shows the band diagram of a subthreshold MOSFET, with definitions for the variables used in this section. I define the drain-to-channel potential, Φ_{dc}, as $V_d - \Psi$, where Ψ is the channel potential and V_d is the voltage at the drain. I define d as the width of the drain-to-channel depletion region. Let δE be the difference of the electron energy from $E_1(z)$:

$$\delta E = E(z) - E_1(z). \tag{2.19}$$

Let $\delta\zeta$ be the difference in the electron momentum angle from $\zeta_1(z)$:

$$\delta\zeta = \zeta - \zeta_1(z). \tag{2.20}$$

2.2.1 Average Electron Energy and Momentum Direction

This subsection determines the average electron energy and the average direction of the momentum as a function of z. To do so, we must solve (2.18) in the new energy and momentum frame of reference of $(z, \delta E, \delta\zeta)$ as defined by (2.19) and (2.20). We can perform this transformation through a Jacobian matrix transformation of (2.18) into this new coordinate system. The transformation is

$$\begin{aligned}
\begin{bmatrix} \frac{\partial f}{\partial E} \\ \frac{\partial f}{\partial \zeta} \\ \frac{\partial f}{\partial z} \end{bmatrix} &= \begin{bmatrix} \frac{\partial \delta E}{\partial E} & \frac{\partial \delta\zeta}{\partial E} & \frac{\partial z}{\partial E} \\ \frac{\partial \delta E}{\partial \zeta} & \frac{\partial \delta\zeta}{\partial \zeta} & \frac{\partial z}{\partial \zeta} \\ \frac{\partial \delta E}{\partial z} & \frac{\partial \delta\zeta}{\partial z} & \frac{\partial z}{\partial z} \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial \delta E} \\ \frac{\partial f}{\partial \delta\zeta} \\ \frac{\partial f}{\partial z} \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{dE_1(z)}{dz} & -\frac{d\zeta_1(z)}{dz} & 1 \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial \delta E} \\ \frac{\partial f}{\partial \delta\zeta} \\ \frac{\partial f}{\partial z} \end{bmatrix}.
\end{aligned} \tag{2.21}$$

After the coordinate transformation, the operators on the right-hand side of (2.18) become:

$$\left(q\mathcal{E}(z) - \frac{E_R}{\zeta\lambda}\right)\frac{\partial f}{\partial E} \rightarrow \left(q\mathcal{E}(z) - \frac{E_R}{\lambda} - \frac{dE_1(z)}{dz}\right)\frac{\partial f}{\partial \delta E}, \tag{2.22}$$

$$q\mathcal{E}\frac{1-\zeta^2}{\zeta E}\frac{\partial f}{\partial \zeta} \rightarrow \left(q\mathcal{E}\frac{1-\zeta^2}{\zeta E} - \frac{d\zeta_1(z)}{dz}\right)\frac{\partial f}{\partial \delta\zeta}. \tag{2.23}$$

I will choose $E_1(z)$, $\zeta_1(z)$ such that the right-hand side of these two terms equal zero. With the appropriate initial conditions, choosing the functions $E_1(z)$, $\zeta_1(z)$ in this way yields reasonable approximations of the average electron energy rigorously defined as the integral of the distribution function over all energies and momentum angles. We can eliminate the dependence on the δE partial derivative in (2.22) by setting $E_1(z)$ as

$$\frac{dE_1(z)}{dz} = q\mathcal{E}(z) - \frac{E_R}{\lambda\zeta}, \tag{2.24}$$

and we can eliminate dependence on the $\delta\zeta$ partial derivative in (2.23) by setting $\zeta_1(z)$ as

$$\frac{d\zeta_1(z)}{dz} = q\mathcal{E}(z)\frac{1-\zeta^2}{\zeta E}. \tag{2.25}$$

The approach followed so far in solving the transport equation is known as "the method of characteristics" and is described in [69, 70]; it is also similar to the numerical method presented in [71]. The functions $E_1(z)$, $\zeta_1(z)$ are called the *characteristics* since they are the flow lines for the hyperbolic P.D.E. operator on the left-hand side.

This model assumes a boundary condition that the electron energy is above the conduction band, that is that the electron has positive kinetic energy. The characteristic functions are only valid in the region inside the boundary conditions. If we assume that the electron has no initial energy, then (2.24) is not valid until its right-hand side is greater than zero since the electron cannot have negative energy in the conduction band. Therefore, I consider this analysis valid when the electrons are accelerated by a sufficient electric field to be 'freed' from the low energies around the conduction band, as seen in Fig. 2.6. I define the breakaway field, $\mathcal{E}_{crit}$ as the minimum electric field at which the electron gains energy at the same rate as it loses

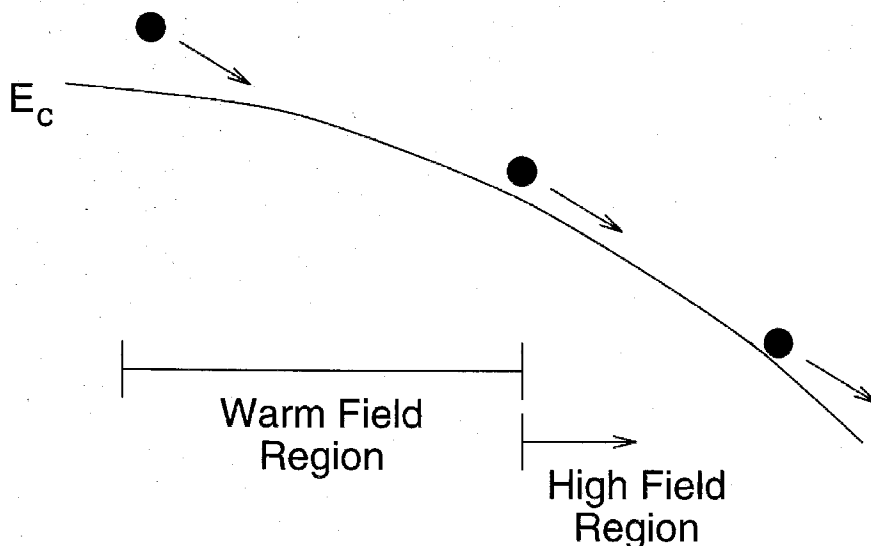


Figure 2.6: Illustration of the electron characteristic paths at different electric fields. In the warm field region, the restoring force due to collisions is larger than the force due to the electric field. In the high field region, the force due to the electric field is larger than the restoring force due to collisions. When an electron reaches an electric field larger than the restoring force due to collisions it can gain significant energies above the conduction band.

energy to phonon collisions. The breakaway field is expressed as $\frac{E_R}{q\lambda}$, which for our parameters is 9.7V/μm. This result is consistent with *streaming* behavior obtained for spatially-uniform electric fields of 10V/μm from Monte Carlo studies [59]; the resulting steady-state average electron energy was 250meV, which is consistent with my approximation. I define the distance from the channel end of the depletion region to this breakaway field as z_{crit}, as illustrated in Fig. 2.5. The average electron analysis will be applied to the region beyond z_{crit}. The physics of the electrons leaving the conduction band is analogous to ballistic transport in the sense that the electrons do not make enough phonon collisions to restore the electrons back to the conduction band [72, 73]. Typically ballistic transport is considered in semiconductor devices where device dimensions are on the order of the mean free path length [72, 73].

For this discussion, I approximate a constant number of electrons per unit length leaving the conduction band throughout the depletion region after z_{crit}. In a more accurate model, the number of electrons per unit length will increase initially due to the larger window of ζ that the electrons can escape the conduction band, and

decrease eventually because at some point a majority of electrons will have left the conduction band. Also, before the electrons reach the breakaway field, the electron distribution function will be a function of the transport in the 'warm' field region. The effects of these phenomena will be the subject of another study.

We now return to the solution for the characteristic equations (2.24) and (2.25). The solution for $\zeta_1(z)$ from (2.25) is (derived in Appendix A.1)

$$\left(1-(\zeta_1+\delta\zeta)^2\right)\left(\frac{E_1(z)+\delta E}{E_a}\right)\left(\frac{1-(\zeta_1+\delta\zeta)}{1+(\zeta_1+\delta\zeta)}\right)^{\frac{E_R}{q\mathcal{E}\lambda}}=1, \qquad (2.26)$$

where E_a is a parameter dependent upon the initial conditions. When $E_1(z)+\delta E$ becomes large, $\zeta_1(z)+\Delta\zeta$ goes to one; therefore once the electrons reach higher energies they are more likely to be pointed along the field direction. Since the field direction has the smallest required electric field to free an electron from the conduction band, the electrons also start directed along the field direction. Because of these two facts, I can safely approximate ζ as near one for energies above the conduction band. In Section 2.4, I will consider the distribution function away from $\zeta = 1$.

Assuming that $\zeta \approx 1$, $E_1(z)$ can be obtained by integrating (2.24) in the region from z_{crit} to z. The solution for $E_1(z)$ is

$$E_1(z) = qV(z) - qV(z_{crit}) - E_R\frac{z-z_{crit}}{\lambda}, \qquad (2.27)$$

as performed in Appendix A. The energy the electron gains after reaching z_{crit} is the difference of the potential from z_{crit} to drain and the number of phonon collisions made in the distance $z - z_{crit}$.

2.2.2 Solution of the Distribution function

This section obtains solutions for the distribution function, first for the simpler case of low energies, and then for all energies. These solutions are obtained by solving the differential equations along the average electron trajectory. I begin with the

Boltzmann Transport Equation for $\zeta \approx 1$ given in terms of E:

$$\frac{\partial f}{\partial z} + \left(q\mathcal{E} - \frac{E_R}{\lambda}\right)\frac{\partial f}{\partial E} = F(T)\frac{E_R^2}{2\lambda}\frac{\partial^2 f}{\partial E^2} - \frac{f}{L(E)}; \tag{2.28}$$

I will make the transformation to the ΔE coordinate system later in this section. I will assume that $1/L(E)$ has a negligible effect on the solution for sufficiently low energies, because $L(E)$ changes rapidly for energies between 1eV and 2eV. For electron energies where $1/L(E)$ is negligible, (2.28) reduces to the following lossless partial differential equation:

$$\frac{\partial f}{\partial z} + \left(q\mathcal{E} - \frac{E_R}{\lambda}\right)\frac{\partial f}{\partial E} = F(T)\frac{E_R^2}{2\lambda}\frac{\partial^2 f}{\partial E^2}. \tag{2.29}$$

For energies where $1/L(E)$ is not negligible, we will factor out the solution of (2.29) from (2.28). I accomplish the factorization by defining $f(z,E)$ in terms of two functions

$$f(z,E) = g(z,E)a(z,E). \tag{2.30}$$

I define $g(z,E)$ as the solution to (2.29); therefore the solutions of the lossless region continue to apply into this region, but are simply attenuated by the function $a(z,E)$ due to impact ionization. Since this procedure is valid for all values of $1/L(E)$, I require that $a(z,E) = 1$ everywhere for the lossless region. The attenuation function $a(z,E)$ can be substituted in (2.28), which results in the following partial differential equation for $a(z,E)$ in terms of z and E (please see detailed derivation in Appendix A.2.1):

$$\frac{\partial a}{\partial z} + \left(q\mathcal{E} - \frac{E_R}{\lambda} - F(T)\frac{E_R^2}{\lambda}\frac{\frac{\partial g}{\partial E}}{g}\right)\frac{\partial a}{\partial E} = F(T)\frac{E_R^2}{2\lambda}\frac{\partial^2 a}{\partial E^2} - L(E)a. \tag{2.31}$$

If electrons are uniformly generated in z, as in the classical MOS capacitor experiments [52], then the Boltzmann transport equation which models this phenomena is (2.31). This measured data followed a lucky-electron model; I also observe a lucky-electron model response from our device in this configuration.

Let us consider the solution of (2.29) by first changing from E to δE coordinates

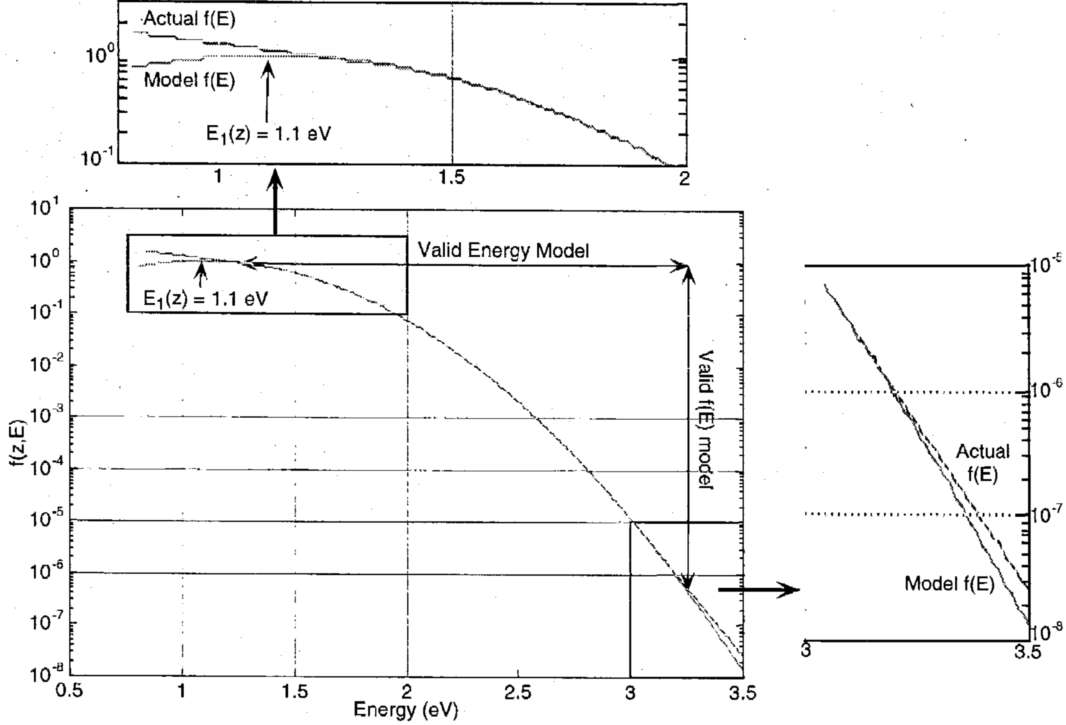


Figure 2.7: Picture of the distribution function for an electron in the drain-to-channel beyond $z = z_{crit}$. This figure compares my Gaussian approximation model versus the solution using the exact boundary condition. For large positive energies, the distribution function does not change as fast as the Gaussian, but rather at a slope around kT_c.

to get

$$\frac{\partial f}{\partial z} = F(T)\frac{E_R^2}{2\lambda}\frac{\partial^2 f}{\partial \delta E^2}, \tag{2.32}$$

which is equivalent to the well-known diffusion equation in space and energy coordinates. To solve (2.32), we need to determine the boundary and initial conditions. The first boundary condition is determined by noting that the solution of (2.32) is bounded in the energy and space coordinates. To obtain additional boundary/initial conditions we assume that the electron leaving at z_{crit} dominates the behavior of hot-electron injection and impact-ionization for a wide range of drain voltages. We can write the solution to (2.32) with this approximation as

$$f(z, \delta E) = \exp\left(-\frac{\lambda}{z - z_{crit}}\left(\frac{\delta E}{2E_R}\right)^2\right), \tag{2.33}$$

which when converted back to E coordinates yields

$$f(z,E) = \exp\left(-\frac{\lambda}{z - z_{crit}}\left(\frac{E - E_1(z)}{2E_R}\right)^2\right). \tag{2.34}$$

The distribution function, f, diffuses from its initial state in energy δE as it evolves in z. Note that the width of the diffusion solution increases as the square root of the position, and therefore the number of phonon collisions.

Figure 2.7 shows the resulting distribution function for the complete model with all three boundary conditions and for the approximate Gaussian model of an impulse at $z = z_{crit}$ for $E = 0$. The Gaussian approximation has two limitations:

1. If the electron distribution function at z_{crit} is approximated by a Boltzmann distribution with an 'effective' temperature, T_c, due to the warm-field transport, then the following equation gives an additional initial condition

$$f(z = z_{crit}, E = \delta E) = e^{-\frac{E}{kT_c}}. \tag{2.35}$$

2. If I model a constant number of electrons per unit length leaving the conduction band, the resulting second boundary condition is

$$f(z, E = 0) = H(z - z_{crit}). \tag{2.36}$$

Including these two effects increases greatly the complexity of an analytic solution.

I now proceed to simplify (2.31) and solve for the attenuation function $a(z,E)$. Since L(E) is extremely small until the electron energy is larger than the bandgap, the energy boundary condition is nearly at the silicon bandgap. One might expect that the second-order derivative in E is small since a constant solution is as diffuse as possible. I show in Appendix A.2.2, that the $F(T)\frac{E_R^2}{\lambda}\frac{\partial^2 g}{\partial E^2}$ term is a second-order perturbation in $\frac{E_R}{q\mathcal{E}(d)l_{ion}}$, and the $\frac{\partial g}{\partial E}/g$ term is a small quantity in the ranges of interest. For this device, the region of significant impact ionization collisions corresponds to the region of the largest electric fields; therefore $q\mathcal{E}\lambda$ will be much larger than E_R.

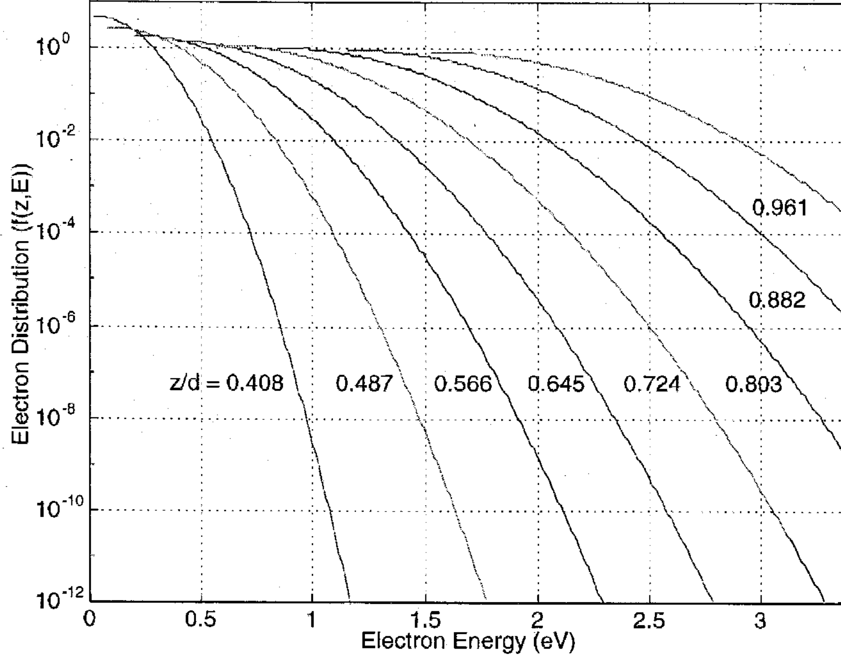


Figure 2.8: Plot of $f(z,E)$ as a function of energy (E) for several locations in the drain-to-channel depletion region for $\Phi_{dc} = 4.0$V, $\lambda = 6.5$nm, and $N_a = 10^{17}\text{cm}^{-3}$. For these parameters, $z_{crit} = 0.36d$. I normalized $f(z,E)$ over energy; normalizing over position would simply change the global scale factor. The width of the drain-to-channel depletion region is d.

By ignoring these terms, I reduce the problem to the form

$$\frac{\partial a}{\partial z} + \left(q\mathcal{E} - \frac{E_R}{\lambda}\right)\frac{\partial a}{\partial E} = -L(E)a. \tag{2.37}$$

with a constant boundary condition at energies near the bandgap.

To obtain a simple form solution for (2.37), I make the following two approximations. First, if L(E) has no spatial or electric field dependence, then with $a(z, E = 0) = 1$, there is no gradient in z. Second, if $q\mathcal{E}(d) - \frac{E_R}{\lambda}$ is nearly constant at the drain edge, then the distribution function is nearly uniform as a function of position. This approximation is good since most of the impact-ionization occurs near the drain edge, and our measurements are sensitive to the effects at the drain edge. Therefore, I assume that the partial derivative of a with respect to z is zero. With these

approximations, we can solve (2.31) by inspection:

$$a(z,E) = \exp\left(-\frac{1}{(q\mathcal{E}(z)\lambda - E_R)}\int_{E=0}^{E}\frac{\lambda}{L(E)}dE\right). \tag{2.38}$$

The solution to $f(z,E)$ can be written as

$$f(z,E) = \exp\left(-\frac{\lambda}{z - z_{crit}}\left(\frac{E - E_1(z)}{2E_R}\right)^2\right)a(z,E). \tag{2.39}$$

Figure 2.8 shows a plot of the above distribution function versus electron energy at several positions in the drain-to-channel depletion.

2.3 Comparing Theory with Experiment

This section compares experimental data for impact ionization and hot electron injection measured in MOSFETs to model calculations that employ the derived distribution functions of the previous sections. More specifically I relate the theoretical distribution functions with the measurable quantities of substrate current and gate currents at the device terminals.

In the next five subsections, I will show measured data of hot-electron injection and impact ionization, and derive device models to relate our data to our model of the electron distribution function. Section 2.3.1 describes the structure and operation of the MOSFET used to measure hot-electron-injection and impact-ionization currents. The depletion region is characterized by its Φ_{dc} and its doping profile; therefore, Section 2.3.2 presents experimental data on the Early voltage to show that the channel doping profile can be approximated by a step junction. Section 2.3.3 derives the distribution function at the drain edge of the depletion region, because the hot-electron injection and impact ionization currents are primarily functions of the distribution at the drain edge. Section 2.3.4 presents the device model of hot-electron-injection current and show the supporting measured data. Section 2.3.5 presents the device model of impact-ionization current and show the supporting measured data.

2.3.1 Device Structure and Bias Condition

Figure 2.1 shows the cross-section of the n-type MOSFET that I use to measure hot-electron injection and impact ionization. I place the MOSFETs in a moderately-doped ($1 \times 10^{17}\text{cm}^{-3}$) substrate to achieve a high threshold voltage. These devices are fabricated through the MOSIS foundry; the moderately doped substrate is formed by the $p$base implant layer, normally employed to form the base of an npn bipolar transistor. The $n$well is used to isolate the $p$base layer from the substrate, to facilitate the measurement of the substrate current due to the electron multiplication. The hot-electron-injection properties of the devices are identical with or without the isolating $n$well.

Since I operate MOSFET transistors in subthreshold, the transport in the drain-to-channel depletion layer is the high-field region of interest. In the subthreshold operating region of a MOSFET, the surface potential in the channel is fixed [74], and thus the high-field phenomena does not depend on the channel current. The electrons in the drain-to-channel region are accelerated to high energies by the high electric fields in the drain-to-channel depletion region. Subthreshold current levels are sufficiently small as to not affect the surrounding electrostatics and the electron density in the channel is low which allows us to ignore the electron-electron scattering effects.

The moderately doped substrate helps the measurements in two ways. First, the high threshold voltage allows hot-electron injection in subthreshold by guaranteeing that the floating-gate to drain voltage always stays positive, and thus electrons reaching the silicon–silicon-dioxide barrier are swept to the floating-gate. Second, the higher substrate doping yields a higher hot-electron-injection and impact-ionization efficiency for a given drain-to-source voltage. The high electric fields responsible for the hot-electron injection and impact ionization in this subthreshold MOSFET structure are consistent with measurements in deep submicron FETs.

I deduce the impact-ionization current by measuring the substrate current, and the hot-electron-injection current from measurements of the gate current. I compute

Φ_{dc} from the sum of the drain-to-source voltage and the source-to-channel potential (Φ_{sc}). In a subthreshold MOSFET, source-to-channel potential is a direct function of the source current (I_s):

$$\Phi_{sc} = kT \ln\left(\frac{I_s}{I_{th}}\right) + 150\text{mV}, \tag{2.40}$$

where I_{th} is the current at threshold, and $\Phi_{sc} = 150$mV at threshold. The measured substrate and gate currents are proportional to the transistor's channel current for identical drain-to-channel and floating-gate-to-drain voltages, as confirmed by the experimental measurements in this section, and in [35].

I define impact-ionization and hot-electron-injection current efficiency as the probability that an electron starting in the channel will be involved either in an impact ionization event or will be injected in the gate, respectively. My model presently assumes that the differences in the initial and final density of states result in small perturbations to the overall solution for this calculation.

2.3.2 Drain-to-Channel Depletion Width Dependence on Φ_{dc}

The electrostatics of the drain-to-channel depletion region near the silicon-silicon-dioxide interface are probed through the channel length modulation phenomenon, which is known as the Early effect. The Early effect manifests itself as a weak dependence of the channel current on drain to source voltage and is a result of the drain-to-channel depletion region width changes with applied Φ_{dc}. I analytically model this behavior as

$$I_{source} = \frac{I_{sat}}{1 - \frac{d(\Phi_{dc})}{L}}. \tag{2.41}$$

Assuming the step-junction model in (2.48), we can model the source current as

$$I_{source} = \frac{I_{sat}}{1 - \frac{\sqrt{\frac{2\epsilon_{si}}{qN_A}}}{L}\sqrt{\Phi_{dc}}}, \tag{2.42}$$

where I_{sat} is the saturation current of the transistors at small drain-channel voltages, ϵ_{si} is the permittivity of silicon, q is the charge of an electron, and N_A is the substrate doping. We can determine the dependence of Φ_{dc} on d by measuring the FET channel current for different drain-to-source voltages. I generalize the original Early-voltage (V_o) definition as

$$V_o = I_{ds} / \frac{\partial I_{ds}}{\partial V_{ds}}, \tag{2.43}$$

where V_{ds} is the drain voltage. For a given saturation current, Φ_{sc} is nearly constant; therefore, I approximate $d\Phi_{dc} = dV_{ds}$ and I define

$$V_o(\Phi_{dc}) = I_{ds} / \frac{\partial I_{ds}}{\partial \Phi_{dc}}. \tag{2.44}$$

From our definition of V_o and (2.42), the Early voltage for a step junction is

$$V_o(\Phi_{dc}) = \sqrt{\frac{2qN_aL^2}{\epsilon}}\sqrt{\Phi_{dc}} - 2\Phi_{dc}. \tag{2.45}$$

If the drain diffusion is not a step junction, but follows the relation $d(\Phi_{dc}) = AL\Phi_{dc}^x$, then the source current is related as

$$I_{source} = \frac{I_{sat}}{1 - A\Phi_{dc}^x}, \tag{2.46}$$

and the resulting V_o is

$$V_o(\Phi_{dc}) = \frac{1}{xA}\Phi_{dc}^{(1-x)} - \frac{1}{x}\Phi_{dc}. \tag{2.47}$$

Figure 2.9 shows the measured early voltage versus Φ_{dc} for two values of gate voltage. The slope of both curves plotted in Fig. 2.9 are in close agreement with step junction behavior. However, the drain-to-channel electric field, and therefore the effective doping, changes significantly with gate voltage. For gate voltages below 7V, the junction behavior is a weak function of gate voltage and roughly corresponds to a substrate doping of 9.53 $\times 10^{16}$cm^{-3}. The effective junction doping steadily increases for larger gate voltages; from Fig. 2.9, a gate voltage of 9.72V corresponds to an effective substrate doping of 2 $\times 10^{18}$cm^{-3}. The dependence of the channel length

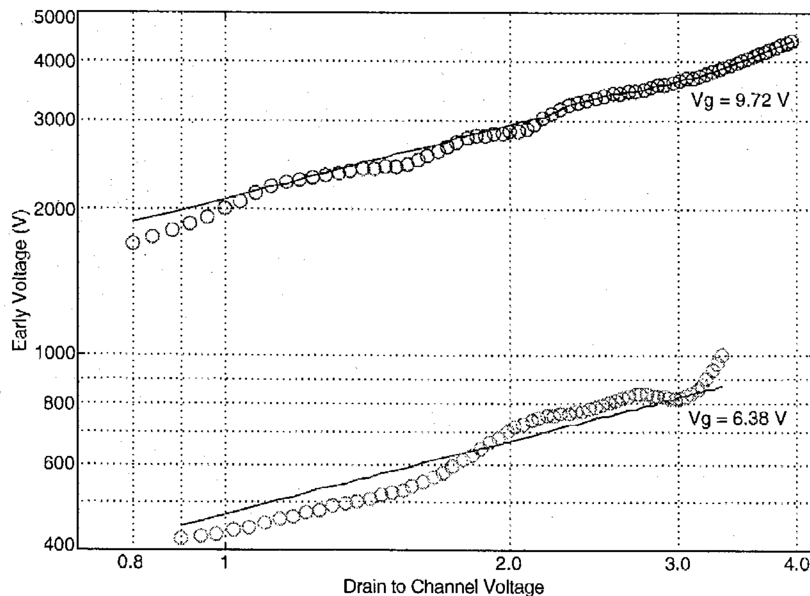


Figure 2.9: Measured Early voltage versus Φ_{dc} for two values of gate voltage. This log-log plot brings out the power law of the junction, and therefore the effective doping profile. For a particular gate voltage, the Early voltage curve is consistent with results from a step junction with $L = 16\mu$m and $N_a = 9.53 \times 10^{16}$cm^{-3}. The curves are similar to the lower V_g case up to gate voltages of 7V, and then gradually increases until it reaches the higher V_g curve. For a gate voltage of 9.72eV, the Early voltage curve is consistent with results from a step junction with $N_a = 2 \times 10^{18}$cm^{-3}.

modulation on the gate voltage does not affect the development of impact-ionization or hot-electron-injection models presented in this chapter.

2.3.3 Electron Distribution Function at the Drain Edge

The gate and substrate currents depend on the distribution function at the drain edge of the drain-to-channel depletion region. To calculate the distribution function for a given drain-to-channel potential (Φ_{dc}), I need to compute the average electron energy at the drain edge (E_{dc}). Here I compute the width of the drain-to-channel depletion region (d), the electric field at the drain edge $\mathcal{E}(z = d)$, and z_{crit} as defined in Section 2.2.

The width of the drain-to-channel depletion region is related to the applied drain

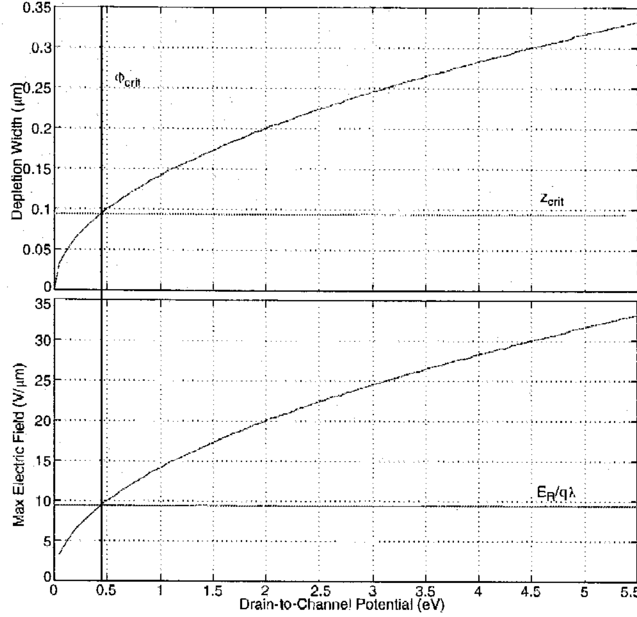


Figure 2.10: Computed parameters for the drain-to-channel depletion region. The plots show the width of the drain-to-channel depletion width, and the maximum electric field in the depletion region (at the drain edge). The substrate doping was $1 \times 10^{17}\text{cm}^{-3}$, and $\lambda = 6.5\text{nm}$. For this device, z_{crit} is approximately one third of the distance into the depletion region for $\Phi_{dc} = 3.6\text{V}$.

to channel potential (Φ_{dc}), and is given by

$$d = \sqrt{\frac{2\epsilon_{si}\Phi_{dc}}{qN_A}}, \tag{2.48}$$

where ϵ_{si} is the permittivity for silicon, q is the charge of an electron, and N_A is the substrate doping level. This relation assumes a step junction doping profile that was justified in the previous section. The electric field, $\mathcal{E}(z)$, is given by

$$\mathcal{E}(z) = \frac{qN_A}{\epsilon}z = \Phi_{dc}\frac{2z}{d^2}, \tag{2.49}$$

and the maximum electric field in the drain-to-channel depletion region, which is the electric field at d, is given by

$$\mathcal{E}(z = d) = \sqrt{\frac{2qN_a\Phi_{dc}}{\epsilon_{si}}}. \tag{2.50}$$

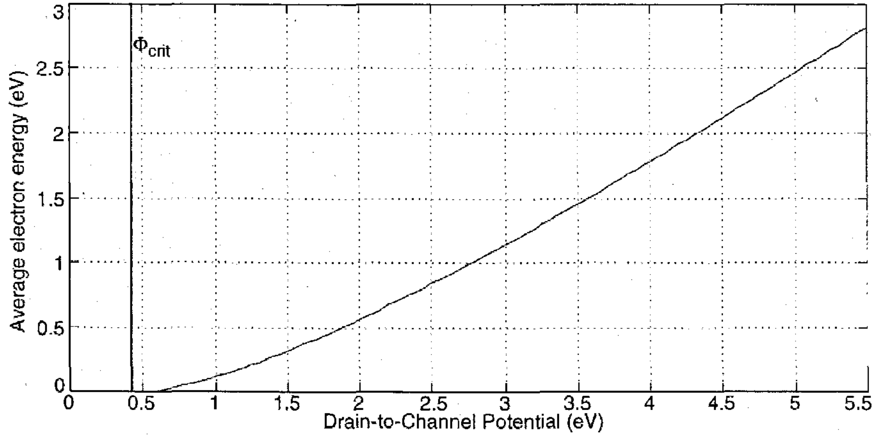


Figure 2.11: Plot of the average electron energy as a function for the drain-to-channel depletion potential. The substrate doping was $1 \times 10^{17}\text{cm}^{-3}$, and $\lambda = 6.5\text{nm}$.

Recall that z_{crit} is the position where the applied electric field is equal to $\frac{E_R}{q\lambda}$; I solve for z_{crit} as

$$z_{crit} = \left(\frac{E_R \frac{d}{2\lambda}}{q\Phi_{dc}}\right) d = \frac{E_R}{\lambda}\frac{\epsilon_{si}}{qN_A}. \tag{2.51}$$

I define Φ_{crit} as the potential at z_{crit} by

$$\Phi_{crit} = \left(\frac{E_R d}{2\lambda}\right)^2 \frac{1}{\Phi_{dc}} = \left(\frac{E_R}{\lambda}\right)^2 \frac{\epsilon_{si}}{2qN_A}. \tag{2.52}$$

Figures 2.10 and 2.11 illustrates these parameters calculated as a function of Φ_{dc} for a substrate doping of $N_A = 1 \times 10^{17} cm^{-3}$. Figure 2.12 shows z_{crit} versus substrate doping.

The average electron energy, E_{dc}, is computed by integrating $\mathcal{E}(z)$ from z_{crit} to the drain edge, d, as in (2.27):

$$E_{dc} = E_1(z = d) = q\Phi_{dc} - E_R\frac{d}{\lambda} + q\Phi_{crit}. \tag{2.53}$$

By substituting d from (2.48), the average electron energy is explicitly a function of Φ_{dc}:

$$\sqrt{E_{dc}} = \sqrt{q\Phi_{dc}} - \sqrt{q\Phi_{crit}}. \tag{2.54}$$

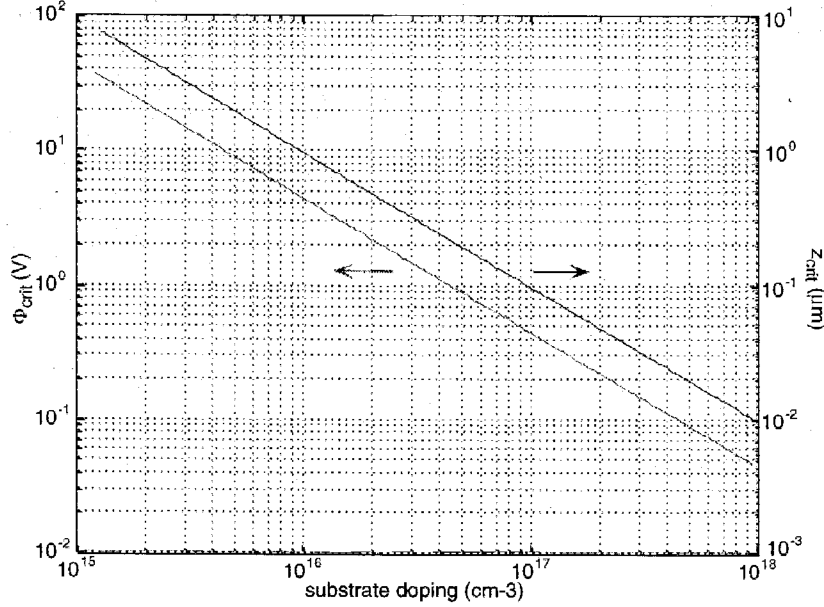


Figure 2.12: Plot of z_{crit} versus MOSFET substrate doping for $\lambda = 6.5$nm.

Figure 2.13 shows the average electron energy versus the drain-to-channel potential for several substrate dopings. The distribution function at the drain edge can be derived from (2.39) as

$$f(d,E) = \exp\left(-\frac{\lambda}{d - z_{crit}}\left(\frac{E - E_{dc}}{2E_R}\right)^2\right) a(d,E),) \tag{2.55}$$

where $a(d,E)$ is

$$a(d,E) = \exp\left(-\frac{1}{(q\mathcal{E}(d)\lambda - E_R)}\int_{E=0}^{E}\frac{\lambda}{L(E)}dE\right). \tag{2.56}$$

Functional form of $L(E)$ in (2.17) results in two distinct $a(d,E)$ regions from the definition in (2.56). For small energies, $|\log(a(d,E))|$ is much less than 1; therefore I approximate $a(d,E)$ by expanding the first exponential in (2.56) as

$$1 - a(d,E) = \frac{\int_0^E \exp\left(\frac{\lambda}{L(E)}dE\right)}{q\mathcal{E}(d)\lambda - E_R}. \tag{2.57}$$

I will use (2.57) to model low electron impact ionization efficiencies for low electron

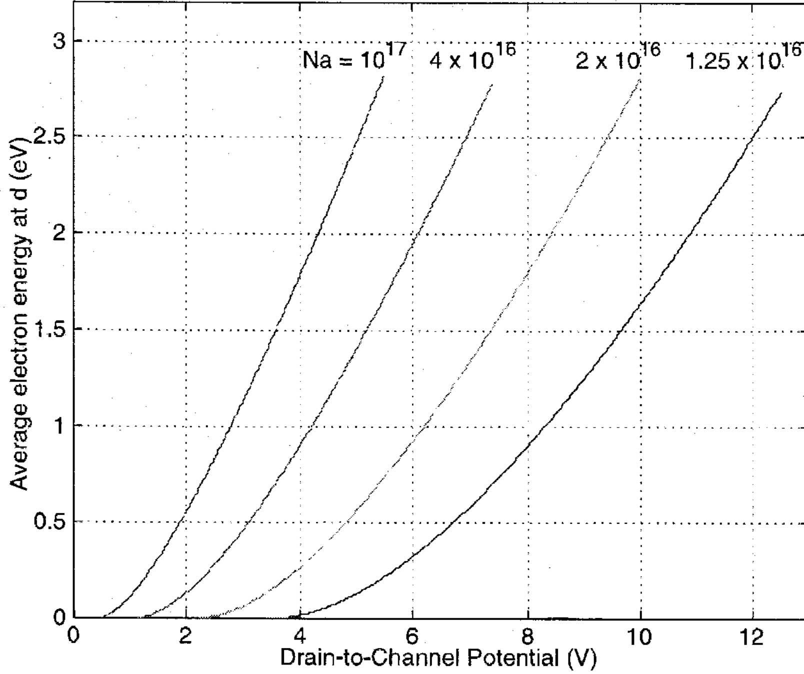


Figure 2.13: Plot of the average electron energy at the drain edge versus drain-to-channel potential for four different MOSFET substrate dopings.

energies, corresponding to I_{ion}/I_s less than 0.2. For larger energies (>2.4 eV), L(E) is a slowly varying function, and can be approximated by expanding the exponential; by expanding around 2.8eV, we can approximate $a(d, E)$ in this region as

$$a(d, E) = \exp\left(-\frac{3.29 \times 10^{-4} \frac{\lambda}{l_{ion}}(E + 3.7eV)}{q\mathcal{E}(d)\lambda - E_R}\right). \tag{2.58}$$

Later in this section, I use (2.58) to model the hot-electron injection current. Figures 2.14 and 2.15 show numerically computed plots for $a(d, E)$ and $1-a(d, E)$ as a function of energy for different values of Φ_{dc}. Figure 2.16 shows a numerically computed plot of the distribution function at d for three values of Φ_{dc}.

2.3.4 Modeling of Hot-Electron Injection in MOS Transistor

This subsection solves for the hot-electron injection (gate current) efficiency. If

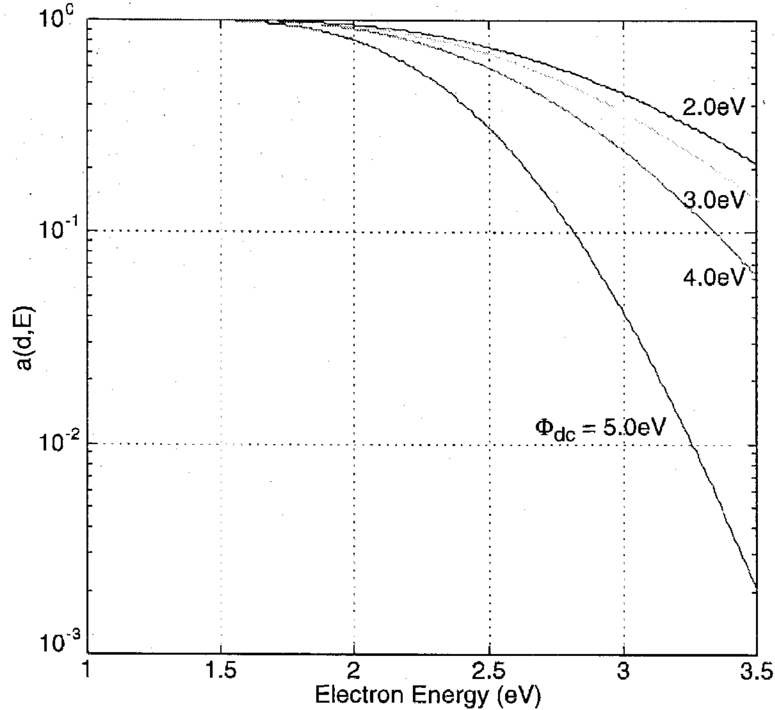


Figure 2.14: Attenuation function at the drain edge due to impact ionization collisions as a function of electron energy for five different Φ_{dc}. I calculated these curves by evaluating (2.56) by numerical integration.

Φ_{dc} and V_{fgd} are fixed, then the injection current should be proportional to the source current, because each electron makes an equal contribution to the injection current. Figure 2.17 shows the measured probability of hot-electron injection versus source current. This data was taken by sweeping the source voltage, and measuring the resulting source current and injection current for a constant floating-gate voltage and drain voltage. The floating-gate–to–drain voltage was explicitly held fixed. For subthreshold biases, the drain-to-channel voltage is fixed throughout the sweep since only the source voltage is adjusted to change the source current. For each sweep in Fig. 2.17, the injection efficiency is very nearly constant for subthreshold current levels. I obtained different drain-to-channel potentials by applying different drain voltages. For above threshold biases, Φ_{dc} is a function of both the source and drain.

Figure 2.18 shows measured data of hot-electron injection efficiencies as a function of drain-to-channel voltage for three channel currents. The experimental evidence

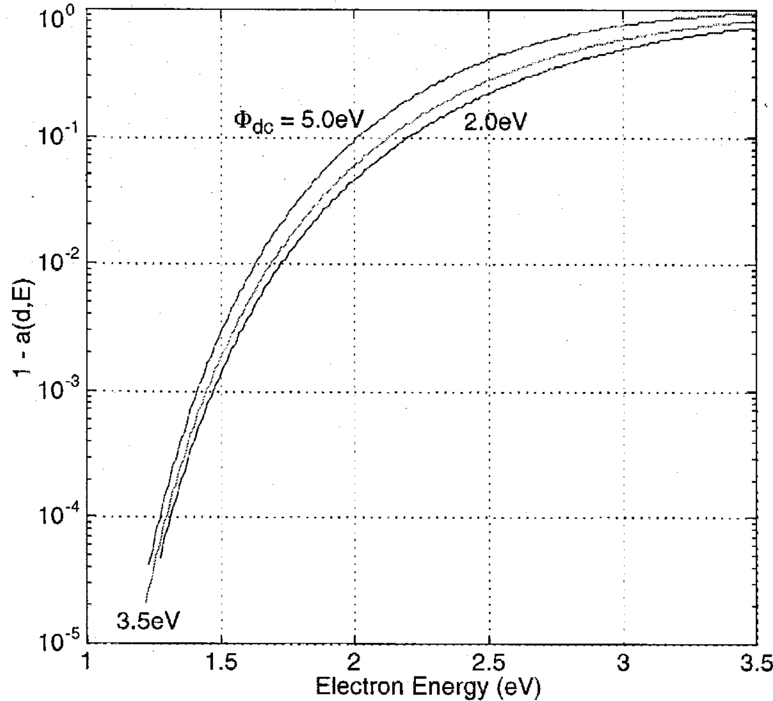


Figure 2.15: 1 - $a(z,E)$ at the drain edge due to impact ionization collisions as a function of electron energy for three different Φ_{dc}. I calculated these curves by evaluating (2.56) by numerical integration.

in Fig. 2.18 shows that the hot-electron-injection efficiency is independent of source current. The electrons go over the barrier by a similar process to thermionic emission, except that the electrons are not described by an equilibrium distribution function. The electrons which have energies higher than the SiO_2 barrier level can participate in hot-electron injection. To investigate hot-electron injection, we need to consider the electron distribution function around the SiO_2 barrier near the drain edge.

I am only considering the case where the floating gate potential is much greater than the drain potential. Two effects are important in this operating region. First, any electrons that can enter the SiO_2 region will be swept towards the floating gate by the resulting electric field. Second, most of the electrons with sufficient energy to surmount the SiO_2 barrier will be at the drain edge. The electron distribution at the SiO_2 barrier energy will be exponentially decreasing as one moves from the drain edge; therefore the hot-electron-injection probability is entirely dependent upon the

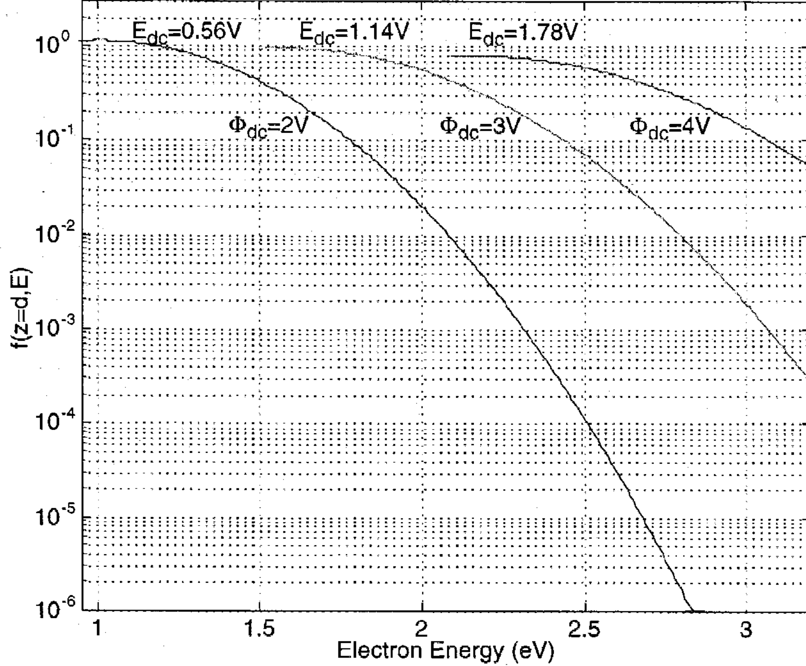


Figure 2.16: Plot of the distribution function at the drain edge as a function of energy for three drain-to-channel potentials. I calculated these curves by evaluating (2.56) by numerical integration.

distribution function at $z = d$. In my experimental setup, the floating gate-to-drain voltage (V_{fgd}) is always positive; therefore, I am certain that most of the electrons are surmounting the barrier at the drain edge. It is well-known that the actual height of the oxide barrier is a function of V_{fgd} due to the barrier lowering effects [90]. I estimate from my data, that the effective barrier height is 2.8eV, and this value is used in my models.

We now solve for the hot-electron injection efficiency. From (2.58), and substituting the values for l_{ion}, λ, and $\mathcal{E}(d)$, the resulting distribution function is

$$f(d, E) \propto \mathrm{e}^{-\frac{\lambda}{d - z_{crit}}\left(\frac{E - q\Phi_{dc} + E_R\frac{d}{\lambda} - q\Phi_{crit}}{2E_R}\right)^2} e^{-\frac{7.102}{\sqrt{E_{dc}}}}, \tag{2.59}$$

where I obtain $a(d, E)$ by expanding around an energy of 2.8eV, as described earlier in this section. Swings in the barrier heights of 100-200meV around 2.8eV add only

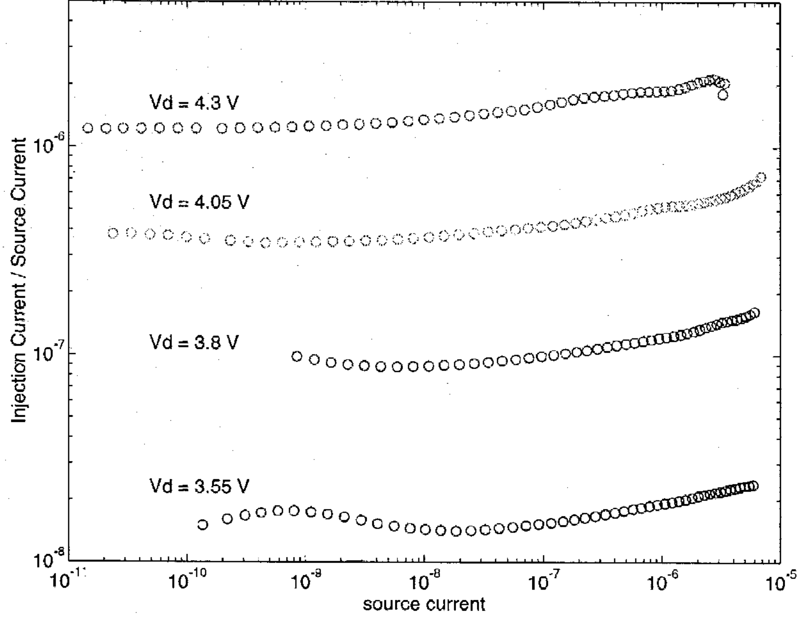


Figure 2.17: Electron injection efficiency versus source current for three values of drain voltage. The curves were obtained by holding the drain and floating-gate fixed while sweeping the source voltage. Electron injection efficiency is defined as the ratio of the injection current and the source current. For subthreshold currents, this ratio is constant for a given drain voltage since the drain-to-channel voltage is constant in this regime.

a small correction term. For electrons to inject into the oxide, they need to have a direction pointing towards the SiO_2; in the discussion, I will show that the angle distribution only weakly affects the hot-electron–injection efficiency.

The hot-electron-injection efficiency (δ) is approximated as proportional to the integral of the distribution function at the drain edge from the oxide barrier energy to the vacuum level, which I write as

$$\delta = \frac{I_{inj}}{I_s} \propto \int_{E=E_{ox}(V_{fgd})}^{E=\infty} f(d,E)dE. \tag{2.60}$$

I approximate this integral to obtain δ as

$$= B_2 e^{-\frac{\lambda}{d-z_{crit}}\left(\frac{E_{ox}(V_{fgd})-\Phi_{dc}+E_R\frac{d}{\lambda}-\Phi_{crit}}{2E_R}\right)^2} e^{-\frac{7.102}{\sqrt{E_{dc}}}}, \tag{2.61}$$

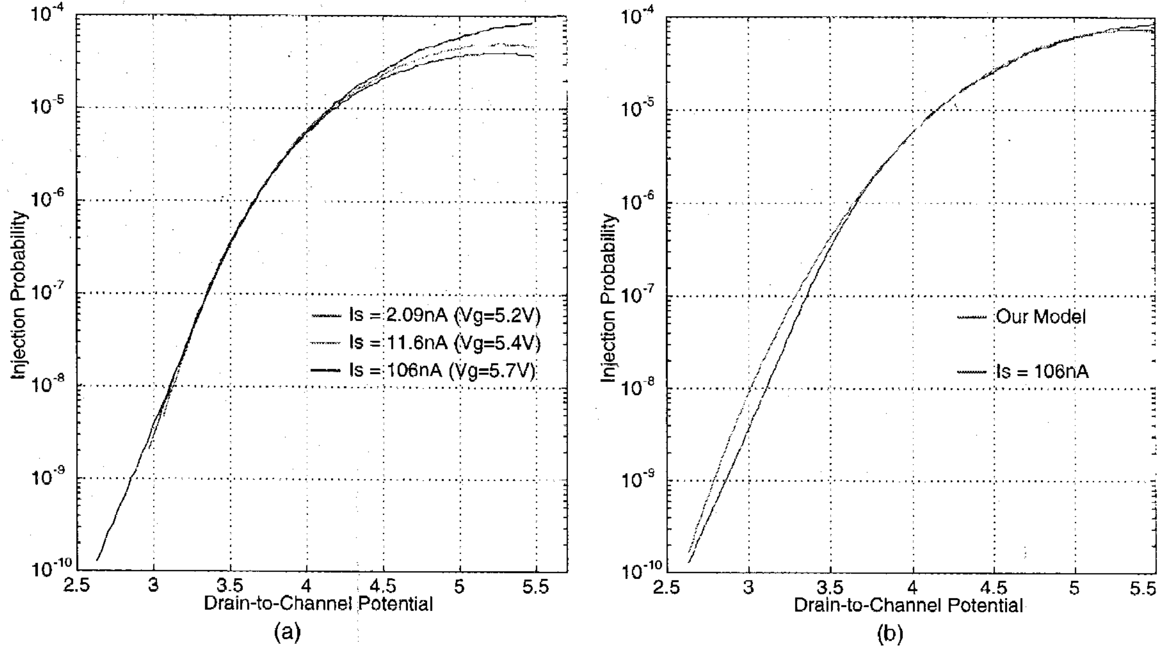


Figure 2.18: (a) Measurements of hot-electron-injection efficiency versus drain-to-channel voltage for several values of source current. I simultaneously measured the substrate and gate currents for different drain-to-channel voltages which gives us the impact-ionization and hot-electron-injection efficiency. The drain-to-channel voltage is computed from the source current and the drain-to-source voltage. For each sweep, I used a constant gate voltage to choose a particular channel current; the actual oxide barrier height changes slightly due to image force lowering, because the floating-gate–to–drain voltage is not constant. E_{ox} will roughly change by 100meV as predicted by image force lowering. (b) Measurement of hot-electron-injection efficiency versus Φ_{dc} compared with a curve fit to the analytic model in (2.61).

where $B_2 = 4.55 \times 10^{-3}$ is a measured constant of proportionality which includes the effects of the electron angles. In this expression, I explicitly show that oxide barrier energy at the drain edge, $E_{ox}(V_{fgd})$ is a function of V_{fgd}. Figure 2.18 shows (2.61) fitted to the injection efficiency data. The curve fit shows close agreement to (2.61) except at Φ_{dc} greater than 5.3V, since average-electron energy is near the energy of the silicon–silicon-dioxide barrier. The significant deviation at low Φ_{dc} is probably due to ignoring the band-structure effects.

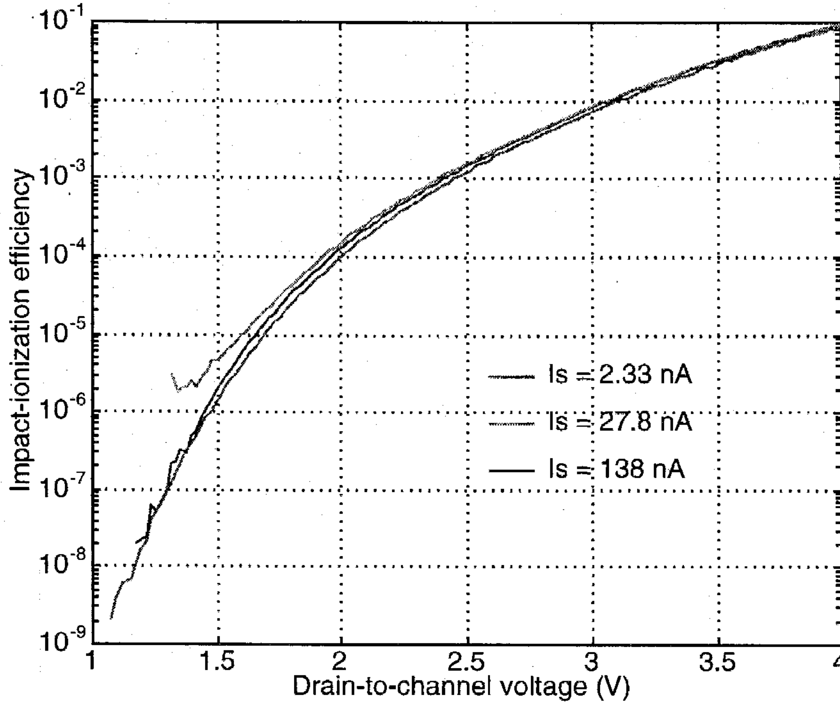


Figure 2.19: Measurements of impact-ionization efficiency vs. drain to channel voltage for three source currents (gate voltages).

2.3.5 Electron Impact Ionization

This subsection solves for the impact ionization (substrate current) efficiency (α), which is the ratio of the substrate current (I_{ion}) to the current at the source (I_s) of the transistor. Figure 2.19 shows measured data of impact-ionization efficiencies as a function of drain-to-channel voltage for several currents. The impact ionization efficiency $\alpha(E_{dc})$ can be thought of as the probability of an electron starting in the channel undergoing an impact ionization in the drain-to channel depletion region.

I calculate $\alpha(E_{dc})$ by computing the ratio of the integral of the distribution function of the electrons undergoing an impact-ionization collision to the integral of the starting distribution function:

$$\alpha(E_{dc}) = \frac{I_{ion}}{I_s} = \frac{\int_{-\infty}^{\infty} (f(0,E) - f(d,E))\, dE}{\int_0^{\infty} f(0,E) dE}. \tag{2.62}$$

To evaluate $f(0,E)$, I substitute $\int_0^\infty f(d,E)dE$ for $\int_0^\infty f(0,E)dE$ for the case of no impact ionization ($a(d,E)=1$ for all energies); this mathematical operation explicitly shows the influence of $1-a(d,E)$ on $\alpha(E_{dc})$. In the region where the electron distribution function can be approximated by a Gaussian (see (2.33)), the equation for $\alpha(E_{dc})$ becomes

$$\alpha(E_{dc}) = \frac{\int_0^\infty (1-a(d,E))\, e^{-\frac{\lambda}{d-z_{crit}}\left(\frac{E-E_{dc}}{2E_R}\right)^2} dE}{\int_0^\infty e^{-\frac{\lambda}{d-z_{crit}}\left(\frac{E-E_{dc}}{2E_R}\right)^2} dE}. \tag{2.63}$$

From (2.63) and the distribution function given by (2.55), we can solve for $\alpha(\Phi_{dc})$. I approximate this solution by expanding the function in the exponent around the function's maximum value in E; the critical energy, E_c, that maximizes this function is nearly equal to the linear function

$$E_c = E_{c1} - \frac{3}{5}\left(1 - \frac{E_{c1}}{2.85\text{eV}}\right)(E_{dc} - 0.95\text{eV}), \tag{2.64}$$

where E_{c1} is

$$E_{c1} = \left(E_R^2\sqrt{\frac{2\epsilon_{si}(119\text{eV})(0.95\text{eV})}{q^2 N_a \lambda^2}}\right)^{2/5}. \tag{2.65}$$

By expanding the terms in the exponent of (2.63) around E_c, using the relation in (2.54), I get the approximate solution of $\alpha(\Phi_{dc})$ as

$$\alpha(\Phi_{dc}) = \exp\left(-\sqrt{\frac{119\text{eV}}{\left(1-\frac{E_{c1}}{2.85\text{eV}}\right)\Phi_{dc}}}\right)$$

$$\exp\left(\frac{\sqrt{\frac{qN_a\lambda^2}{2\epsilon_{si}}}}{\sqrt{\Phi_{dc}}-\sqrt{\Phi_{crit}}}\left(\frac{E_{c1}-\frac{3}{5}\left(1-\frac{E_{c1}}{2.85\text{eV}}\right)(E_{dc}-0.95\text{eV})}{2E_R}\right)^2\right) \tag{2.66}$$

This integration yields a smoothed version of $L(E)$. The factor in the second exponential decreases rapidly for Φ_{dc} greater than 3.5V. Note that we can use (2.63) to compute $a(d,E)$, and therefore can compute $L(E)$ from measurements of α as a function of Φ_{dc}. We calculate the $L(E)$ function plotted in Fig. 2.4 using the mea-

surements of α as a function of Φ_{dc}.

So far, this model only models the impact-ionization current in the drain-to-channel depletion region and not the impact-ionizations in the drain regions due to the high-energy electrons entering this region. In this region, the electron transport changes in two ways. First, the electric field in the drain region is negligible. Second, the electrons start at high energies. The transport in this region is similar to other experimental setups used to measure quantum-yield by other researchers [75], and therefore, provides yet another check of my modeling from additional experimental results.

We will follow the distribution of a single electron on its (x, y, z) path through the drain. I represent the position along this path by s; this approach allows us to concentrate on the energy distribution and neglect the momentum scattering. The following Boltzmann transport equation models the electron distribution in (s, E)

$$\lambda\frac{\partial f}{\partial s} = E_R\frac{\partial f}{\partial E} + F(T)\frac{E_R^2}{2}\frac{\partial^2 f}{\partial E^2} - \frac{\lambda}{L(E)}f. \tag{2.67}$$

Our goal is to solve for the number of electrons that are removed from the distribution over all position; therefore I integrate over all s

$$f(s, E)|_{s=0}^{\infty} = E_R\frac{\partial \bar{f}}{\partial E} + F(T)\frac{E_R^2}{2}\frac{\partial^2 \bar{f}}{\partial E^2} - \frac{\lambda}{L(E)}\bar{f}, \tag{2.68}$$

where I define

$$\bar{f} = \frac{1}{\lambda}\int_0^{\infty} f(s, E)ds. \tag{2.69}$$

Since all the electrons will return to equilibrium energies at the conduction band, $f(\infty, E) = 0$ for all energies, except for $E = 0$. Also, since I am considering a single electron starting at an initial energy, E_i, $f(0, E) = \delta(E - E_i)$, where $\delta(\cdot)$ is the delta function. I simplify (2.68) by modeling the distribution function for low impact-ionization efficiencies, and by assuming a negligible second derivative in energy; both simplifications are consistent with the model level in the drain-to-channel region. The

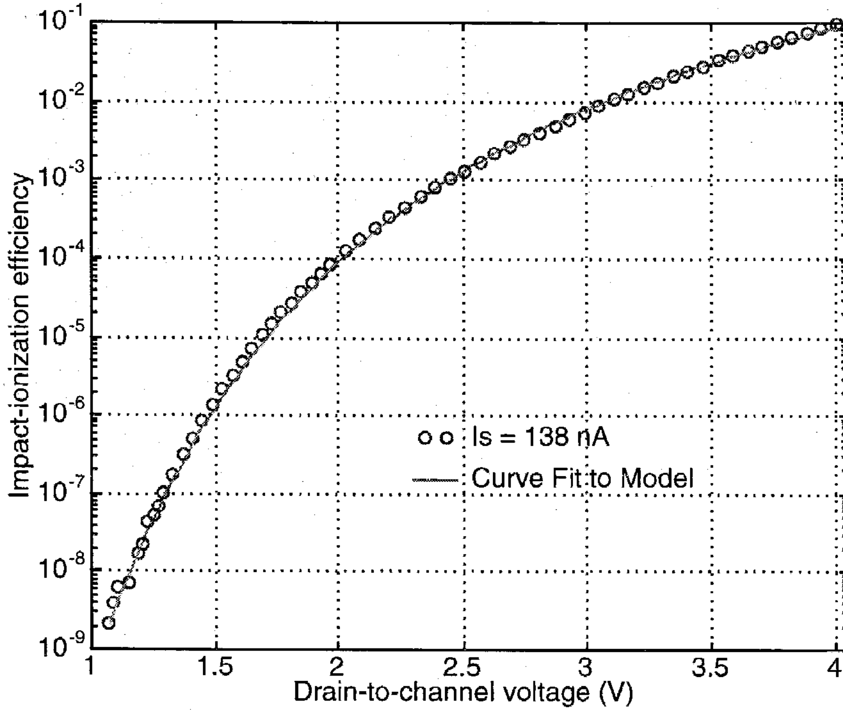


Figure 2.20: Measurement of impact-ionization efficiency versus Φ_{dc} compared with a curve fit to the analytic model in (2.71). The source voltage was held at zero and the gate voltage was held at 5.8V.

solution to this simplified form of (2.68) is

$$1 - \bar{f} = \frac{1}{E_R} \int_0^{\infty} \frac{\lambda}{L(E)} dE. \tag{2.70}$$

Combining the impact-ionization models in the drain and drain-to-channel regions, I model E_{dc} as

$$\alpha(E_{dc}) = \frac{\left(\frac{1}{q\mathcal{E}(d)\lambda - E_R} + \frac{1}{E_R}\right) \int_0^{\infty} \frac{\lambda}{L(E)} e^{-\frac{\lambda}{d - z_{crit}} \left(\frac{E - E_{dc}}{2E_R}\right)^2} dE}{\int_0^{\infty} e^{-\frac{\lambda}{d - z_{crit}} \left(\frac{E - E_{dc}}{2E_R}\right)^2} dE}. \tag{2.71}$$

This modification increases $\alpha(E_{dc})$, and therefore increases the effect of $L(E)$. Figure 2.19 shows experimental measurements of α versus drain-to-channel potential for three different channel currents. Figure 2.20 shows a curve fit (solid line) of (2.71)

and (2.66) to the experimental data with the parameters defined in Section 2.1; the fit closely agrees with the measured data. I measured the impact ionization efficiency by assuming at most only a single impact ionization event will occur for each starting electron, which is reasonable for the low impact-ionization efficiencies of the measurements shown in Fig. 2.20.

2.4 Discussion

In the previous section, I skipped several subtle discussion points for clarity of the presentation. I will address three points in this discussion section. Section 2.4.1 discusses the effect of collision broadening on my derived phonon collision operator. Section 2.4.2 considers the effect of momentum-angle scattering due to optical-phonon collisions. Section 2.4.3 discusses how the electron obtains the necessary angle as well as energy for hot-electron injection.

2.4.1 Effect of Collision Broadening on the Phonon Collision Operator

My model of optical phonon collisions has several interesting extensions. One interesting extension is that the starting point and derivation for acoustical phonons of a particular energy or band to band scattering is similar to my results for optical phonons [60]; therefore with a different set of parameters, (2.12) can model the effects of these three collision processes. It can also be shown when all three processes are taken into account, that the largest contribution comes from the optical phonon collisions.

Second, my model of the optical-phonon collision operator changes only slightly when I model the effect of collision broadening. This result further justifies my use of semi-classical theory, since the τ for optical-phonon collisions for hot-electron injection and impact ionization are at the boundary where Boltzmann transport may not be valid. Collision broadening is due to the limited certainty about a particular energy

between collisions because of the uncertainty principle, $\Delta E\tau = \hbar$, where τ is the average time between collisions. For τ of 10fs, which is the order of the smallest τ, the energy uncertainty is roughly 60meV.

I will now show the effect of collision broadening on my optical-phonon collision operator. I define the broadening function, $B(E)$, as a symmetric function around $E = 0$ that is characterized by a variance of E_d. I formulate the new optical-phonon collision operator as

$$S_{op}(f)\left[\frac{m^*(E)}{c(E)}\right] \approx -\left(e^{\frac{E_R}{kT}}+1\right)f(E)$$

$$\int_{\infty}^{-\infty}\left[B(\Delta E + E_R) + e^{\frac{E_R}{kT}}B(\Delta E - E_R)\right]f(E-\Delta E)d\Delta E. \tag{2.72}$$

This function is a convolution; by taking the Laplace transform, we transform the convolution to a product:

$$\mathcal{L}\left\{S_{op}(f)\left[\frac{m^*(E)}{c(E)}\right]\right\} \approx -\left(e^{\frac{E_R}{kT}}+1\right)f(s_E) + \left[e^{-s_E E_R} + e^{s_E E_R + \frac{E_R}{kT}}\right]B(s_E)f(s_E). \tag{2.73}$$

Since $B(E)$ is real, symmetric function, then $B(s_E)$ will also be a real, symmetric function, and therefore I can expand $B(s_E) = 1 - s_E^2 E_d^2 + O(s_E^4)$; E_d is a measure of the width of the energy uncertainty. If we Taylor expand the s_E terms to second order, we get

$$\mathcal{L}\left\{S_{op}(f)\left[\frac{m^*(E)}{c(E)}\right]\right\} \approx s_E E_R\left(e^{\frac{E_R}{kT}}-1\right)f(s_E)$$

$$+s_E^2\frac{E_R^2}{2}\left(e^{\frac{E_R}{kT}}+1\right)f(s_E) - s_E^2\frac{E_d^2}{2}\left(e^{\frac{E_R}{kT}}+1\right)f(s_E). \tag{2.74}$$

By taking the inverse Laplace transform of this approximate equation, we get the new collision operator as

$$S_{op}(f)\left[\frac{m^*(E)}{c(E)}\right] \approx F(T)\frac{E_R^2 - E_d^2}{2\lambda}\frac{\partial^2 f}{\partial E^2} + \frac{E_R}{2\lambda\zeta}\frac{\partial f}{\partial E}. \tag{2.75}$$

Collision broadening reduces the second derivative in energy term in in the collision

operator expression; Monte Carlo simulations have seen that collision broadening increases the number of high-energy electrons [59] by decreasing the restoring force of the optical phonons. The effect of collision broadening will decrease the higher order derivative terms; the higher order derivatives are more significant at low energies and fields due to the sharper distribution functions. The effect of collision broadening will also be larger at lower energies since $\left(1+\frac{E_R}{E}\right)^{1/2}$ will deviate from 1.

2.4.2 Effect of Momentum Scattering Optical Phonons

In Section 2.1, I assumed that the phonon collisions did not affect the overall electron momentum distribution function. In my previous analysis, I have assumed that the phonon distribution is at its equilibrium level; a careful modeling must also take these effects into account. Phonons have momentum, and the total momentum involved for a phonon absorption or emission must be conserved. My previous phonon-collision operator defined in (2.10) left the distribution function unaffected in the ζ direction; this assumption is one extreme of a continuum of plausible situations. The other extreme assumes that phonons have uniform momentum angle distributions, and therefore a phonon collision will uniformly scatter the electron distribution through all angles. To precisely model this effect, one would need to know the distribution function of momentum for the phonons in the drain–to channel depletion region; this subsection investigates the assumption that a phonon collision uniformly randomizes the momentum distribution. I model the new phonon collision operator by modifying the $f(E+E_R,\zeta)$ and $f(E+E_R,\zeta)$ terms in (2.10) as

$$f(E,\zeta) \to \frac{1}{2}\int_0^1 \left(f(E,\zeta_1)+f(z,E,-\zeta_1)\right)d\zeta_1. \tag{2.76}$$

Again, I will assume that $E \gg E_R$, and therefore $\left(1+\frac{E_R}{E}\right)^{1/2} \approx 1$. The resulting new phonon operator makes an analytical solution of the transport equations extremely difficult. This subsection will consider the resulting behavior when phonon collisions uniformly randomize the momentum distribution function.

One might wonder if this collision operator would result in a nearly isotropic distri-

bution function. The largest electron energy is approximately equal to my previously derived average electron energy for the same z_{crit}. Hot electron injection will occur at low Φ_{dc} only if λ is substantially longer than I defined in Section 2.1; a longer λ requires fewer electron collisions in the drain–to–channel depletion region. The average electron energy for this isotropic model is nearly equal to the energy gain for the $\zeta = 0$ electron. The average electron energy of an electron starting at z and arriving at $z + \lambda$ is approximately described by

$$E_1(z + \lambda) = \sqrt{E_1(z)^2 + (q\mathcal{E}(z)\lambda)^2} - E_R. \tag{2.77}$$

This approximation overestimates the average-electron energy because electrons reaching $z+\lambda$ from z may take several paths that require more phonon collisions, resulting in fewer electrons at higher energies. As in Section 2.2, an electron must gain more energy than is lost by the phonons to leave the conduction band. We define this threshold when $E_1(z_{crit} + \lambda) = E_R$ for $E_1(z_{crit}) = E_R$; we can solve for z_{crit} as

$$q\mathcal{E}(z_{crit}) = \sqrt{3}E_R. \tag{2.78}$$

This z_{crit} is larger than I derived in Section 2.2. The average electron energy will reach a steady state when $E_1(z + \lambda) = E_1(z)$; the steady-state value of $E_1(z)$ is

$$E_1(z) = \frac{(q\mathcal{E}(z)\lambda)^2 - E_R^2}{E_R}, \tag{2.79}$$

and at $z = d$, the steady-state value is

$$E_1(z) = E_R\left(\frac{\Phi_{dc}}{\Phi_{crit}} - 1\right). \tag{2.80}$$

For the parameters given in Section 2.3, $E_1(z) \approx 600$meV for Φ_{dc} =5.5eV; λ would need to increase by more than a factor of two to match the value in Section 2.3. Also, the numerical calculations of impact ionization and hot-electron injection [76] do not give this type of distribution. Therefore it seems highly improbable that

the electron distribution function remains nearly isotropic. We could argue that the relevant electrons have no collisions, as in the physical description of the lucky electron model; this hypothesis could explain the longer λ used in lucky-electron formulations. Since this model does not simultaneously fit both hot-electron-injection and impact-ioniziation currents, this hypothesis also does not satisfy the data; this hypothesis would also require the unlikely possibility that λ is a strong function of Φ_{dc}.

The above arguments show that an isotropic solution is inconsistent with the experimental data; therefore the distribution function in ζ must have bias along the field direction. Consider the case where the distribution function is near to my previous solution assuming $\zeta = 1$. If we are interested in the solution in a neighborhood near $\zeta = 1$, we can Taylor expand the distribution function in the integrals around $\zeta = 1$:

$$S_{op}(f)\left[\frac{m^*(E)}{c(E)}\right] \approx F(T)\frac{E_R^2}{2\lambda}\frac{\partial^2 f}{\partial E^2} + \frac{E_R}{\lambda}\frac{\partial f}{\partial E} - F(T)\frac{\zeta}{\lambda}\frac{\partial f}{\partial \zeta} + \frac{F(T)}{\lambda}\left(\frac{1}{6} + \frac{\zeta}{2} + \zeta^2\right)\frac{\partial^2 f}{\partial \zeta^2} \tag{2.81}$$

where $F(T)$ was defined in (2.13), which is nearly equal to one at room temperature ($T = 300K$). This expansion is valid to third order in partial derivatives, and is very nearly equal to the original expression for energies greater than a few E_R. The first order term moves electrons further towards $\zeta = 1$; this new phonon collision operator modifies the characteristic equation for $\zeta_1(z)$ in (2.25) as

$$\frac{d\zeta_1(z)}{dz} = q\mathcal{E}(z)\frac{1-\zeta^2}{\zeta E} + \frac{1}{\lambda}, \tag{2.82}$$

resulting in $\zeta_1(z)$ moving faster towards 1 than before and being nearly independent of electron energy or electric field. The resulting partial differential equation is

$$\frac{\partial f}{\partial z} + \frac{1}{\lambda}\frac{\partial f}{\partial \zeta} = F(T)\frac{E_R^2}{2\lambda\zeta}\frac{\partial^2 f}{\partial \delta E^2} + \frac{F(T)}{\lambda}\left(\frac{1}{6} + \frac{\zeta}{2} + \zeta^2\right)\frac{\partial^2 f}{\partial \zeta^2}. \tag{2.83}$$

This characteristic equation further illustrates that electrons near $\zeta = 1$ will not quickly move away from $\zeta = 1$. The diffusive second-order term is consistent with a phonon collision uniformly randomizing the electron's momentum angle.

We will now solve the distribution function around $\zeta = 1$. By transforming (2.29)

with the new collision operator in (2.81) to δE coordinates, I approximately get

$$\frac{\partial f}{\partial z}+\frac{1}{\lambda}\frac{\partial f}{\partial \zeta}=F(T)\frac{E_R^2}{2\lambda\zeta}\frac{\partial^2 f}{\partial \delta E^2}+\frac{F(T)}{\lambda}\frac{\frac{1}{6}+\frac{\zeta}{2}+\zeta^2}{\zeta}\frac{\partial^2 f}{\partial \zeta^2}. \tag{2.84}$$

For purposes of analysis, I assume that the solution of (2.84) near $\zeta = 1$ is close to the solution of (2.84) when evaluating the coefficients at $\zeta = 1$:

$$\lambda\frac{\partial f}{\partial z}+\frac{\partial f}{\partial \zeta}=F(T)\frac{E_R^2}{2}\frac{\partial f}{\partial \delta E}+\frac{5}{3}F(T)\frac{\partial^2 f}{\partial \zeta^2}. \tag{2.85}$$

The solution in $(z, \delta E, \zeta)$ to this simplified equation for an impulse at z_{crit} is

$$f=\frac{(\zeta-1)\sqrt{4E_R^2\frac{z-z_{crit}}{\lambda}}}{12\lambda\sqrt{\pi z^3}}\exp\left(-(\frac{9A^2}{16}+\frac{3}{2})\frac{z-z_{crit}}{\lambda}\right)$$
$$\exp\left(\frac{3A}{4}(\zeta-1)-\frac{\lambda}{z-z_{crit}}\left(\frac{3}{8}(\zeta-1)^2+\frac{\delta E^2}{2E_R^2}\right)\right). \tag{2.86}$$

By taking the integral of this solution over all ζ, the solution for $\int_{\zeta=-1}^{1} f(z,E,\zeta)$ is identical to my previous solution for f assuming that $\zeta = 1$.

I will make some qualitative points about the distribution function by extrapolating beyond this perturbative analysis, that is when ζ is significantly smaller than 1. First, the distribution function is diffused more in δE. Second, the average electron energy will be lower along a given $\zeta_1(z)$ path than for the $\zeta = 1$ case we previously analyzed. Third, a lower ζ will result in a slower diffusion of the distribution function away from $\zeta = 1$; the diffusion virtually stops near $\zeta = 0$. Therefore, it is reasonable to believe that the $\zeta = 1$ behavior will dominate the phenomena for most of my range of measurements, but behavior for ζ much smaller than 1 will be significant near and below $E_1(z)$, and this boundary coincides where the Gaussian model breaks down as we previously illustrated in Fig. 2.7. The best approach to justify this model will be with comparisons to full bandstructure Monte Carlo calculations similar to [76], for this device.

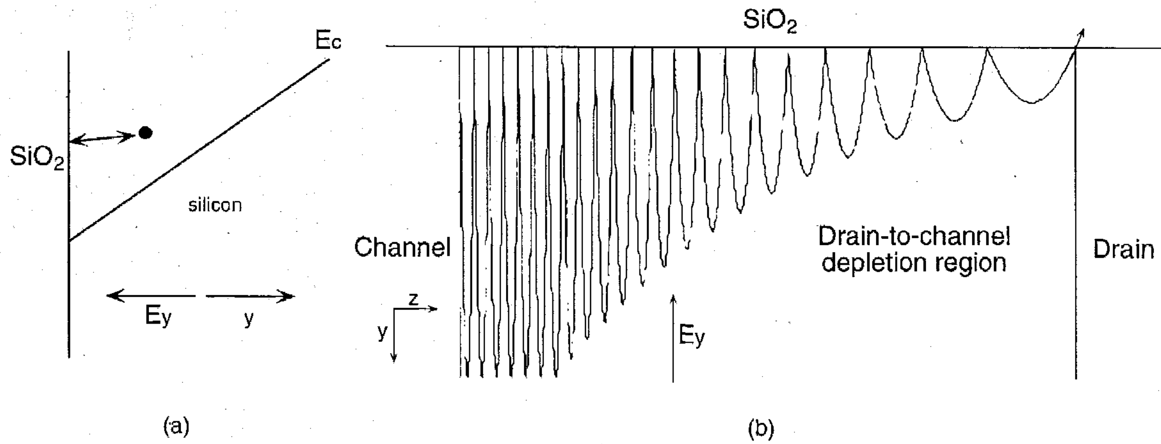


Figure 2.21: Illustration of the 2D nature of the electron transport in the drain-to-channel region. The electron elastically bumps along the barrier at the SiO_2 interface and then gets redirected back towards the interface due to the electric field. In this way, some of the electrons have the proper direction to enter the SiO_2 and eventually reach the floating gate.

2.4.3 How the Electron Gets the Necessary Angle for Impact Ionization

For an electron to reach the floating gate, it must have energy greater than the oxide barrier height and must be directed towards the SiO_2 when the electron reaches that energy. In the previous sections, I only considered how the electrons reach the energy to surmount the SiO_2 barrier; this section addresses how the electrons enter the SiO_2. An electron gains energy due to the electric field in the z direction; electrons are confined by the electric field in the y direction. In the y direction, the electron is being accelerated by the electric field towards the silicon–silicon-dioxide interface, only to be elastically reflected off of this interface back into the silicon. Therefore, the electron's position oscillates in y, and that results in a roughly uniform distribution function in y that is independent from the distribution function in z. Figure 2.21 illustrates the 2D path that the electron takes in the drain-to-channel depletion region. Previous theories by [52, 53] propose that the electrons become directed into the SiO_2 by an elastic ion collision. According to my model, the elastic-ion collisions have little effect in the y direction because randomizing the electron momentum cannot further diffuse

an already uniform distribution function.

Chapter 3 Single-Transistor Synapses

The first step in building VLSI chips that adapt and learn is to develop a silicon analog for a synapse. Synapses are known to play a key role for learning and adaptation in biological systems [83]. I and my colleagues, Diorio and Minch, have successfully developed such a synapse using only a single transistor. We have designed, fabricated, characterized, and modeled arrays of single-transistor synapses. This chapter presents two single-transistor synapses that simultaneously perform long-term weight storage, compute the product of the input and the weight value, and update the weight value according to a Hebbian or a backpropagation learning rule. This combination of functions has not been achieved previously with floating-gate devices, and we believe that this circuit is the first instance of a single-transistor learning synapse fabricated in a standard process.

3.1 Overview of Single-Transistor Learning Synapses

A silicon synapse must perform two computations. First, it must compute the product of the input multiplied by the synapse strength, or the weight of the synapse. Second, it must compute the weight-update rule. For a Hebbian synapse, the weight change is a time average of the product of the input and output activity. In many supervised algorithms such as backpropagation, this weight change is a time average of the product of the input and some fed-back error signal. Both weight-update rules are similar in function.

Figure 3.1 shows the cross sections for the nFET and pFET single-transistor synapses. We build the nFET synapse in a moderately doped ($1 \times 10^{17}\text{cm}^{-3}$) substrate in order to achieve a high threshold voltage. The moderately doped substrate is formed in the 2μm $n$well MOSIS process by the $p$base implant. The pFET synapse FET is the standard pFET in the 2μm $n$well process. Each synapse has its own

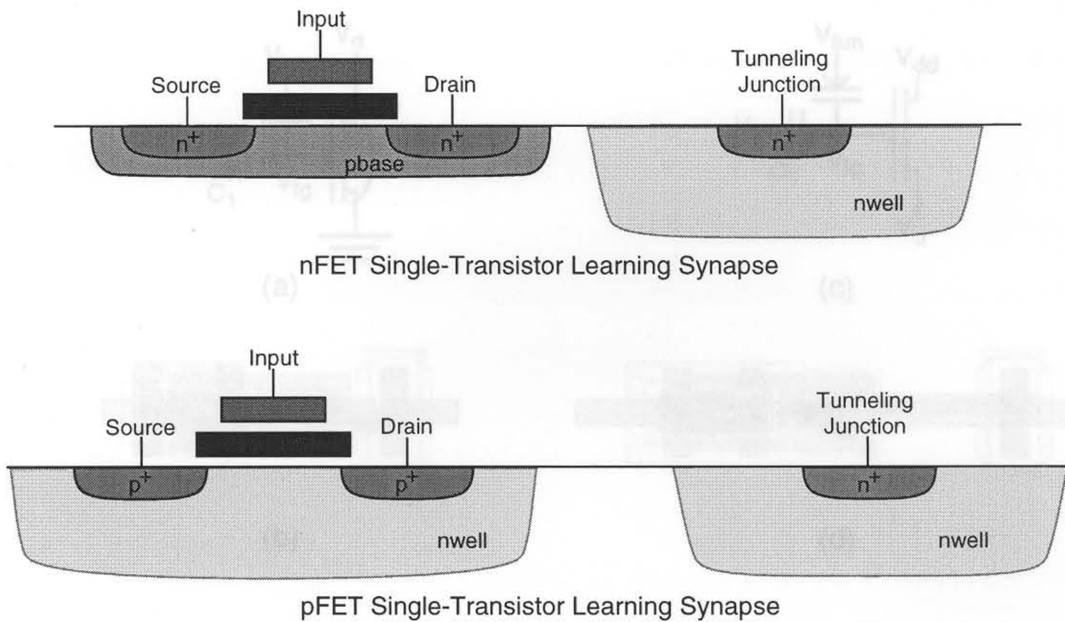


Figure 3.1: Cross section of the nFET and pFET single-transistor synapses in an $n$well MOSIS process. The tunneling junctions used by the single-transistor synapses is a region of gate oxide between the polysilicon floating-gate and $n$well. For the nFET synapse, the pbase implant results in a larger threshold voltage, which results in all the electrons reaching the top of the SiO_2 barrier being swept into the floating gate. The pFET transistor is the standard pFET transistor in the $n$well process.

tunneling junction, which is formed from high-quality gate oxide separating an n-type well region from the floating gate. The tunneling junction removes charge from the floating gate. The particular learning algorithm of an array of synapses depends on the circuitry at the boundaries of the array—in particular the circuitry connected to each of the gate, source, drain, and tunneling lines in a row or column.

The single-transistor synapses possess five necessary properties to build adaptive analog VLSI systems with large synaptic arrays:

1. In the absence of learning, the weight is stored permanently because the floating gate is well insulated by SiO_2.

2. A synapse operating with subthreshold currents computes an output current that is the product of the input signal and the synaptic weight.

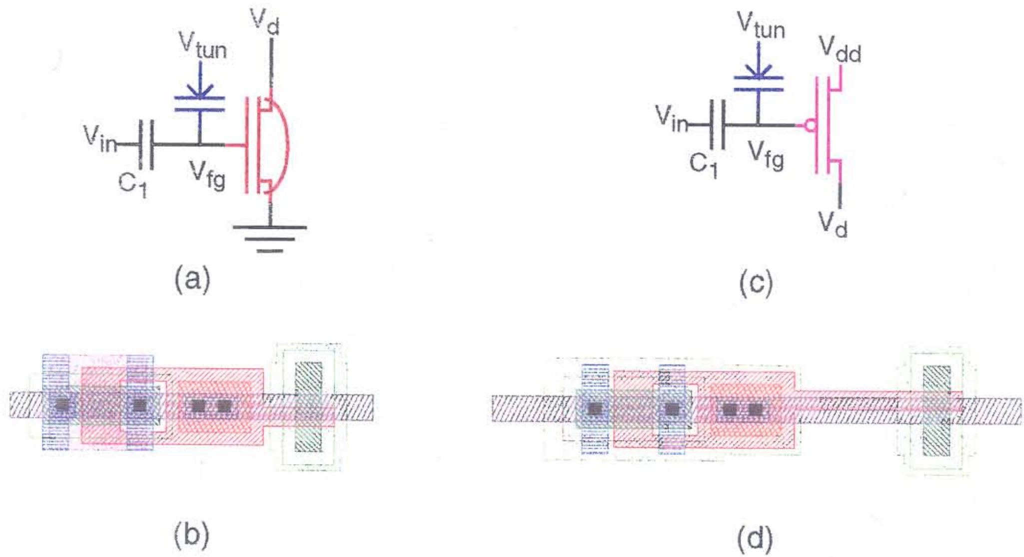


Figure 3.2: Circuit symbols of the nFET and pFET single-transistor synapses. (a) Circuit diagram of the nFET single-transistor synapse with its source connected to ground. (b) Layout of the nFET single-transistor synapse. (c) Circuit diagram of the pFET single-transistor synapse with its source connected to V_{dd}. (d) Layout of the pFET single-transistor synapse.

3. This synapse employs a single transistor, thereby consuming minimal silicon area, and maximizing the number of synapses in a given area.

4. The synapse operating with subthreshold currents dissipates a minimal amount of power; therefore the synaptic array is not power constrained.

5. This synapse can modify its weight using to the floating-gate charge according to Hebbian or backpropagation type learning rules, depending on how various error signals are fed back to the floating gate.

The possible dynamics of different circuit configurations are presented in Chapter 4.

The single-transistor learning synapses use a combination of electron tunneling and hot-electron injection to adapt the charge on the floating gate, and thereby the weight of the synapse. Hot-electron injection adds electrons to the floating gate,

thereby decreasing the weight. Injection occurs for large drain voltages; therefore we can reduce the floating-gate charge during normal feedforward operation by raising the drain voltage. Electron tunneling removes electrons from the floating gate, thereby increasing the weight. The tunneling line controls the tunneling current; thus we can increase the floating-gate charge during normal feedforward operation by raising the tunneling line voltage. The tunneling rate is modulated by both the input voltage and the charge on the floating gate.

3.2 Nonadaptive Behavior

Figure 3.2 shows the circuit models for the nFET and pFET single-transistor synapses. Because the input signals are capacitively coupled to the floating gate, we model voltage and current swings around the circuit's steady-state values. We consider the single-transistor synapse operating with subthreshold channel currents. Many of the behaviors extend qualitatively to above-threshold operation; the quantitative behaviors do not. I describe the subthreshold nFET or pFET channel current in saturation, I_s, for a change in the FET's floating-gate voltage, ΔV_{fg}, source voltage ΔV_s, and drain-to-source voltage, ΔV_{ds}, around a bias current, I_{so}, as [82]

$$\begin{aligned} \text{nFET:} \quad I_s &= I_{so} \exp\left(\tfrac{\kappa_n \Delta V_{fg} - \Delta V_s}{U_T}\right) \exp\left(\tfrac{\Delta V_{ds}}{V_o}\right), \\ \text{pFET:} \quad I_s &= I_{so} \exp\left(\tfrac{\Delta V_s - \kappa_p \Delta V_{fg}}{U_T}\right) \exp\left(-\tfrac{\Delta V_{ds}}{V_o}\right), \end{aligned} \tag{3.1}$$

where κ_p is the fractional change in the pFET surface potential due to a change in ΔV_{fg}, κ_n is the fractional change in the nFET surface potential due to a change in ΔV_{fg}, V_o is the Early voltage of the nFET or pFET, and U_T is the thermal voltage, $\frac{kT}{q}$. I define C_T as the total amount of capacitance connected to the floating gate, and C_2 as the capacitance between floating gate and drain (which, for simplicity, is not explicitly drawn). I assume that all the floating-gate devices are matched. I model the floating-gate behavior by equating the currents at the floating gate:

$$C_T \frac{dV_{fg}}{dt} = C_1 \frac{dV_{in}}{dt} + C_2 \frac{dV_d}{dt} + I_{tun} - I_{inj}, \tag{3.2}$$

where I_{tun} and I_{inj} are the floating-gate currents due to electron tunneling and hot-electron injection, a subject that we discuss in Section 3.3 and 3.4. In (3.1), I use a modified form of the Early voltage expression that is consistent with classical formulations for large V_o, and that more closely models the behavior for small V_o. Chapter 5 shows that the Early voltage decreases at large drain-to-source voltages due to impact ionization in the drain-to-channel depletion region.

To analyze the adaptation behavior in FGMOS circuits, I often decompose our variables into components that change at fast and slow rates. The fast-rate variables represent the rapid changes due to the input signals; the slow-rate variables represent the floating-gate charge (the synapse weights). For sufficiently fast input signals, this decomposition is justified because the total floating-gate charge is only affected by the floating-gate currents. I assume that we can decompose the synapse terminal voltages as sums of fast ($\Delta\hat{V}$) and slow ($\Delta\bar{V}$) quantities as

$$\Delta V_d = \Delta\hat{V}_d + \Delta\bar{V}_d,$$

$$\Delta V_s = \Delta\hat{V}_s + \Delta\bar{V}_s,$$

$$\Delta V_{fg} = \Delta\hat{V}_{fg} + \Delta\bar{V}_{fg}. \tag{3.3}$$

Neglecting Early voltage effects, the source current becomes

$$\begin{aligned} \text{nFET:} \quad & I_s = I_{so} W \exp\left(\frac{\kappa_n \Delta\hat{V}_{fg}}{U_T}\right), \quad W = \exp\left(\frac{\kappa_n \Delta\bar{V}_{fg}}{U_T}\right), \\ \text{pFET:} \quad & I_s = I_{so} W \exp\left(-\frac{\kappa_p \Delta\hat{V}_{fg}}{U_T}\right), \quad W = \exp\left(-\frac{\kappa_p \Delta\bar{V}_{fg}}{U_T}\right), \end{aligned} \tag{3.4}$$

where I define W to be the weight of the synapse. The time derivative of W for the nFET synapse is

$$\frac{d\bar{V}_{fg}}{dt} = \frac{U_T}{\kappa_n W}\frac{dW}{dt} = \frac{U_T}{\kappa_n}\frac{d\log(W)}{dt}; \tag{3.5}$$

later in Figs. 3.9 and 3.10, I plot the derivative of $\log(W)$ versus W to illustrate the dependence of source current on the floating-gate currents. At fast timescales, I approximate the tunneling and injection currents to be negligible; I model the

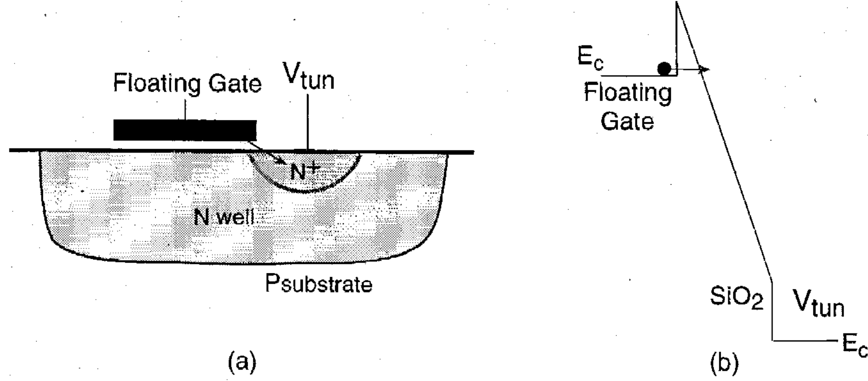


Figure 3.3: Tunneling in an $n$well process. (a) The tunneling junction is the capacitor between the floating gate and the $n$well; we use high-quality gate oxide to reduce the effects of electron trapping. Over a wide range of oxide voltage, most of the tunneling occurs between the floating gate and n^+ diffusion region because this region is accumulated and the higher electric fields at the corner of the floating gate. (b) Band diagram through the tunneling capacitor. For sufficiently large applied electric fields, the electron barrier becomes thin enough that an electron might tunnel through the barrier.

behavior at the fast timescales as

$$C_T \frac{d\hat{V}_{fg}}{dt} = C_1 \frac{d\hat{V}_{in}}{dt} + C_2 \frac{d\hat{V}_d}{dt} \rightarrow \hat{V}_{fg} = \frac{C_1}{C_T}\hat{V}_{in} + \frac{C_2}{C_T}\hat{V}_d. \tag{3.6}$$

By substituting (3.5) for $\frac{d\bar{V}_{fg}}{dt}$ in (3.2), we obtain the dynamical equations for slow timescales as

$$\begin{aligned} \text{nFET:} \quad & \frac{U_T C_T}{\kappa_n W}\frac{dW}{dt} = C_1 \frac{d\bar{V}_{in}}{dt} + C_2 \frac{d\bar{V}_d}{dt} + I_{tun} - I_{inj}, \\ \text{pFET:} \quad & -\frac{U_T C_T}{\kappa_p W}\frac{dW}{dt} = C_1 \frac{d\bar{V}_{in}}{dt} + C_2 \frac{d\bar{V}_d}{dt} + I_{tun} - I_{inj}, \end{aligned} \tag{3.7}$$

where I_{tun} is the electron tunneling current, and I_{inj} is the hot-electron injection current.

3.3 Electron Tunneling

Electron tunneling gives us a method for removing electrons from the floating gate; electron tunneling produces positive floating gate current. Tunneling arises from the fact that an electron wavefunction has finite extent. For a thin enough barrier, this extent is sufficient for an electron to penetrate the barrier. An electric field across the oxide will result in a thinner barrier to the electrons on the floating gate. As illustrated in Fig. 3.3, for a high enough electric field, the electrons can tunnel through the oxide. Increasing the tunneling voltage, V_{tun}, increases the effective electric field across the oxide, which increases the probability of the electron tunneling through the barrier. Typical values for the oxide field range from 0.75V/nm to 1.0V/nm. I will start from the classic model of electron tunneling through a silicon–silicon-dioxide system [78], in which the electron tunneling current is given by

$$I_{tun} = I_0 \exp\left(-\frac{\mathcal{E}_o}{\mathcal{E}_{ox}}\right) = I_0 \exp\left(-\frac{t_{ox}\mathcal{E}_o}{V_{tun} - V_{fg}}\right), \tag{3.8}$$

where $\mathcal{E}_{ox}$ is the oxide electric field, t_{ox} is the oxide thickness, and $\mathcal{E}_o$ is a device parameter that is roughly equal to 25.6V/nm [79]. Figure 3.4 plots tunneling current versus 1/oxide voltage ($V_{tun} - V_{fg}$) showing good agreement with (3.8) in two regions. The cause for two separate regions might be due to tunneling through intermediate traps [80], or due to initially tunneling through the junction edge for low oxide voltages and tunneling through the middle of the junction for high oxide voltages.

When traveling through the oxide, some electrons get trapped in the oxide, which changes the barrier profile. To reduce this trapping effect, I tunneled through high-quality gate oxide, which has far less trapping than interpoly oxide. Each synapse tunneling junction is formed from high-quality gate oxide separating an $n$well region from the floating gate. Both injection and tunneling have very stable and repeatable characteristics. Figure 3.5 shows tunneling current at a fixed oxide voltage versus charge through the oxide; the tunneling current decreases only 50 percent after 12nC of charge has passed through the oxide. This quantity of charge is orders of magnitude

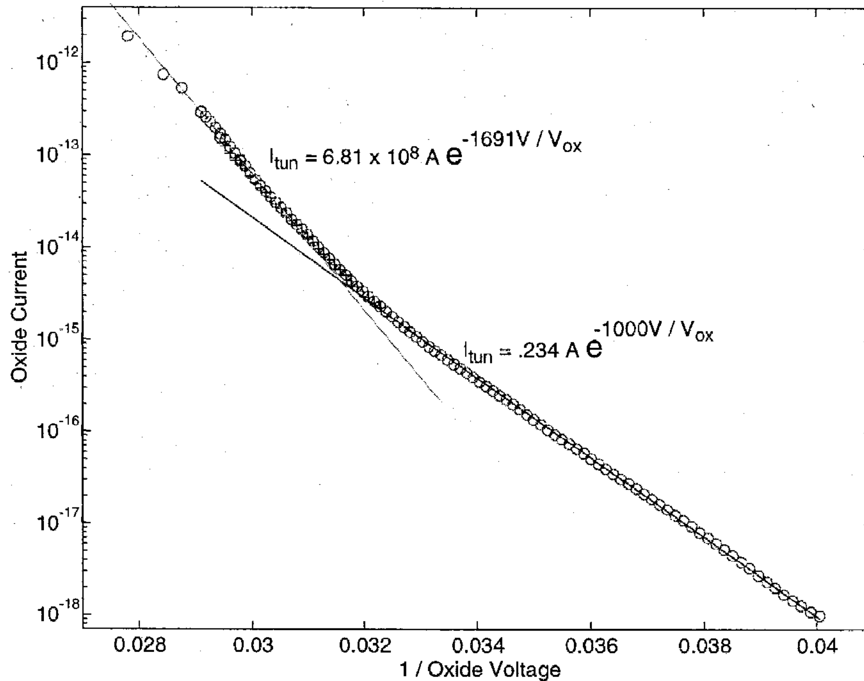


Figure 3.4: Electron tunneling current versus 1/oxide voltage. The two straight line fits are to the classic Fowler-Nordheim expression in (3.8). The two different straight-line regions might be due to tunneling through intermediate traps, or due to initially tunneling through the junction edge for low oxide voltages and tunneling through the middle of the junction for high oxide voltages.

more than we would expect a synapse to experience over a lifetime of operation, because a typical VLSI device will use several orders of magnitude smaller tunneling currents and capacitors.

The next step is to develop a synapse-level model of electron tunneling. Since the floating-gate only moves a few volts in normal operation, expanding V_{fg} as $V_{fg0}+\Delta V_{fg}$ and V_{tun} as $V_{tun0} + \Delta V_{tun}$ in (3.8) results in

$$I_{tun} = I_{tun0} \exp\left(\frac{\Delta V_{tun} - \Delta V_{fg}}{V_x}\right), \tag{3.9}$$

where

$$V_x = \frac{(V_{tun0} - V_{fg0})^2}{t_{ox}\mathcal{E}_o} \tag{3.10}$$

I_{tun0} is the quiescent tunneling current, and ΔV_{tun} is the change in the tunneling

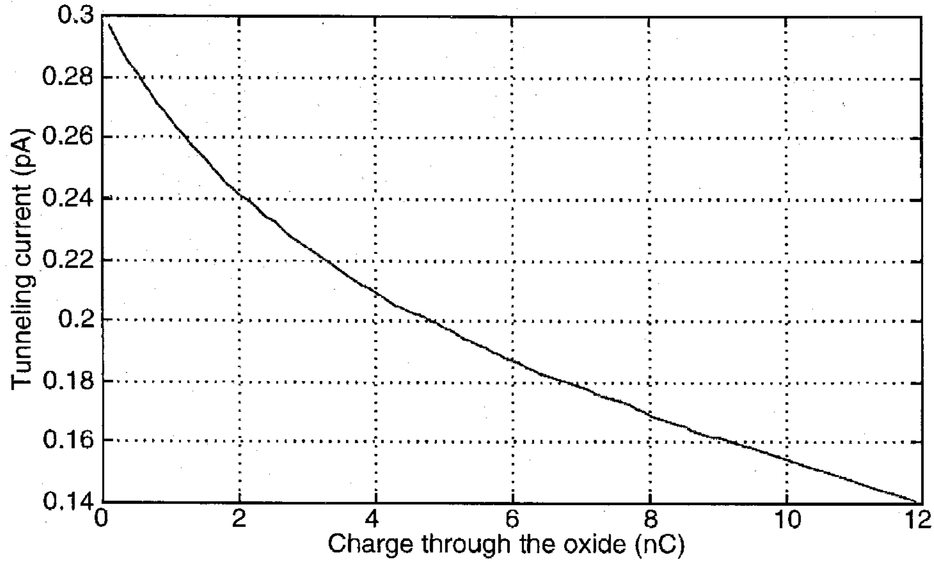


Figure 3.5: Electron tunneling current at a fixed oxide voltage ($V_{ox} = 33$V) versus charge through our tunneling junction. This experiment induced 60V of charge on a 200pF capacitor; to first order, tunneling trap creation is a function of the total charge through the oxide. This measurement characterizes the electron-trapping effect in our tunneling-junction oxide; the degree that the tunneling current decreases shows the effect of increasing electron traps in the oxide.

voltage. For my operating conditions, a typical value of V_x is 1V with the 42nm oxide used in the 2.0μm Orbit process. For a fixed source voltage, I express the tunneling current in terms of this floating-gate transistor's source current by substituting (3.1) into (3.9):

$$n\text{FET} : I_{tun} = I_{tun0} \left(\frac{I_s}{I_{so}} \right)^{-\frac{U_T}{\kappa_n V_x}},$$

$$p\text{FET} : I_{tun} = I_{tun0} \left(\frac{I_s}{I_{so}} \right)^{\frac{U_T}{\kappa_p V_x}}. \tag{3.11}$$

If we fix the synapse terminals such that V_{tun} is high enough for appreciable tunneling, then the weight of this synapse obeys

$$\text{nFET:} \quad \frac{U_T C_T}{\kappa_n I_{tun0}} \frac{dW}{dt} = (W)^{1-\frac{U_T}{\kappa_n V_x}},$$

$$\text{pFET:} \quad -\frac{U_T C_T}{\kappa_p I_{tun0}} \frac{dW}{dt} = (W)^{1+\frac{U_T}{\kappa_p V_x}}. \tag{3.12}$$

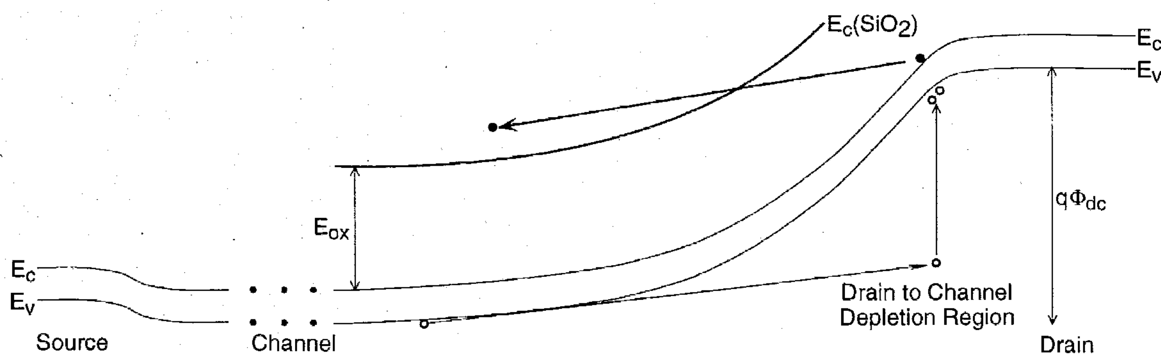


Figure 3.6: Band diagram of a subthreshold pFET transistor under conditions favorable for hot-electron injection. The source of electrons to be injected to the floating-gate is created by hole impact ionization. E_{ox} is the Si–SiO_2 barrier, which is 3.04eV for no field across the oxide.

Figures 3.9 and 3.10 show the plot of $\frac{d}{dt}\log(I_s)$ as a function of I_s for different V_{tun}. The straight line behavior validates the model in (3.12). Starting at an appropriately low (or high) source current (I_s), I measured this data by stepping to the desired drain or tunneling voltage, and letting the tunneling (or injection) current decrease (or increase) the synapse current. The drain voltage was 2V for this tunneling experiment. From (3.5), I plot the derivative of $\log(W)$ versus W to illustrate the dependence of source current on the floating-gate currents. Typical nFET values of $\frac{U_T}{\kappa_n V_x}$ are in the range of $0.1 - 0.3$, while typical pFET values of $\frac{U_T}{\kappa_p V_x}$ are in the range of $0.02 - 0.1$.

3.3.1 Hot-Electron Injection

Hot-electron injection gives us a method to add electrons to the floating gate. The underlying physics of the injection process is to give some electrons enough energy and direction in the drain-to-channel depletion region to surmount the SiO_2 energy barrier, as presented in Chapter 2. To inject an electron onto a floating gate, the MOSFET must have a high-electric-field region to accelerate channel electrons to energies above the silicon–silicon-dioxide barrier, and in that region the oxide electric field must transport the electrons that surmount the barrier to the floating gate. The moderate substrate doping level allows the nFET to easily achieve hot-electron injection in subthreshold operation. The higher substrate doping results in a much

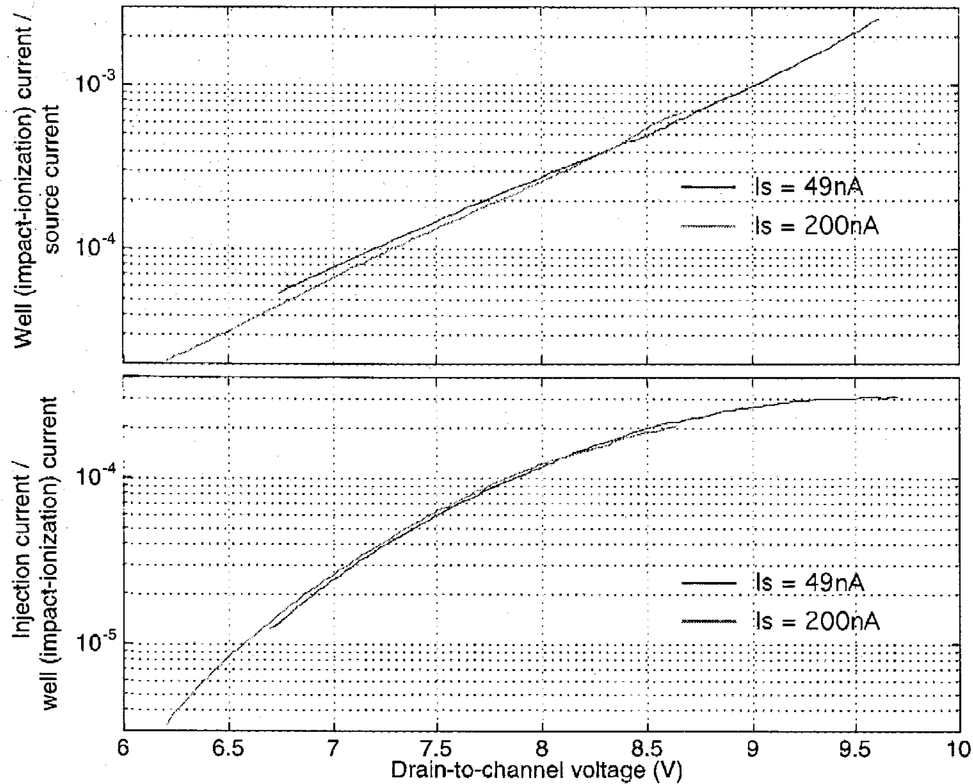


Figure 3.7: Measured data of the current dependences in hot-electron injecting *p*FETs versus the drain-to-channel voltage for two source currents. The top plot shows the ratio of impact-ionization current and source current versus drain-to-channel potential. The lower plot shows the ratio of hot-electron injection and impact-ionization current versus drain-to-channel potential. This ratio saturates around $\Phi_{dc} = 9.0$V due to electrons gaining sufficient energy to surmount the SiO_2 barrier, and gives evidence that hole-impact ionization is the source of the electrons.

higher threshold voltage (6.1V), which guarantees that the field in the oxide at the drain edge of the channel will be in the proper direction for collecting electrons over the useful range of drain voltages. The higher substrate doping results in higher electric fields which yield higher injection efficiencies. The higher injection efficiencies allow the device to have a wide range of drain voltages substantially below the threshold voltage.

One might wonder how *p*FETs, where the current carriers are holes, inject hot electrons onto the floating gate. Figure 3.6 shows the band diagram of a *p*FET operating under bias conditions that are favorable for hot-electron injection. Hot-

hole impact ionization creates electrons at the drain edge of the drain-to-channel depletion region, due to the high electric fields there. These electrons travel back into the channel region, gaining energy as they go. When their kinetic energy exceeds that of the silicon–silicon-dioxide barrier, they can be injected into the oxide and transported to the floating gate. The hole impact-ionization current is proportional to the pFET source current, and is the exponential of a smooth function (f_1) of the drain-to-channel potential (Φ_{dc}). We can express this relationship as follows:

$$I_{impact} = I_p e^{f_1(\Phi_{dc})}, \tag{3.13}$$

where Φ_{dc} is the potential drop from channel to drain. Figure 3.7 plots the ratio of impact-ionization current and source current as a function of Φ_{dc}; this data nearly follows a lucky electron theory and can be analyzed with the methods in Chapter 2. The injection current is proportional to the hole impact-ionization current, and is the exponential of another smooth function (f_2) of the voltage drop from channel to drain. We can express this relationship as follows:

$$I_{inj} = I_{impact} e^{f_2(\Phi_{dc})}. \tag{3.14}$$

Figure 3.7 plots the ratio of hot-electron-injection current versus impact-ionization current as a function of Φ_{dc}; the hot-electron-injection current saturates at Φ_{dc} = 10V due to many hot electrons having enough energy to surmount the SiO_2 barrier. Because the injection current is only a weak function of the floating-gate voltage for a fixed source current (I_p) and Φ_{dc}, I neglect the gate-voltage dependence for this application. A first-principles model of pFET hot-electron injection can be derived using the methods in Chapter 2, and I will publish this material in a later publication.

Figure 3.8 shows the nFET and pFET hot-electron injection efficiency, $\frac{I_{inj}}{I_s}$, as a function of Φ_{dc} for two different source currents. The nFET's substrate implant results in higher electric fields in its drain-to-channel region compared with the electric fields in the pFET; therefore, the nFET's hot-electron–injection efficiency is much

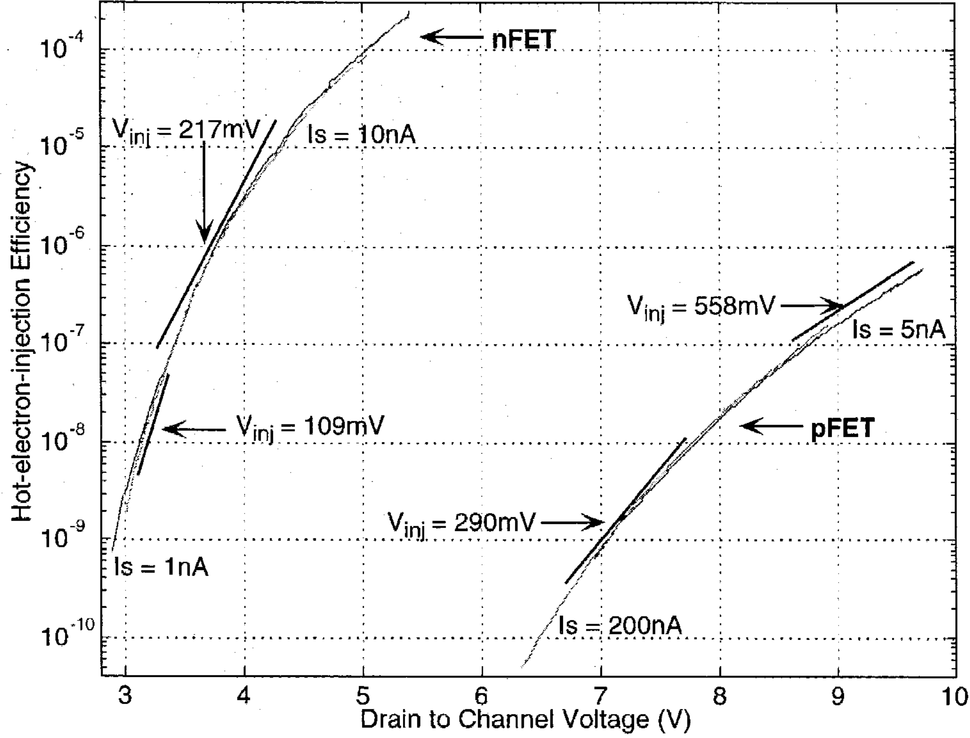


Figure 3.8: nFET and pFET hot-electron injection efficiency ($\frac{I_{inj}}{I_s}$) versus Φ_{dc} for two values of source current. Injection efficiency is the ratio of injection current to source current. The two different source current values are nearly equal, which is consistent with injection efficiency being independent of source current. I show the linearized slope (V_{inj}) on this exponential scale for two Φ_{dc} biases. The slope of both curves on this exponential scale decrease with increasing Φ_{dc}. The moderately doped nFET substrate ($1 \times 10^{17}\text{cm}^{-3}$) increases the efficiency of the nFET hot-electron–injection process by increasing the electric field in the channel.

larger than the pFET's injection efficiency for an equivalent Φ_{dc}. The subthreshold nFET or pFET injection current is proportional to the source current (I_s), and is the exponential of a smooth function (f_3) of the drain-to-channel potential (Φ_{dc}). We can express this relationship as follows:

$$I_{inj} = I_s e^{f_3(\Phi_{dc})}. \tag{3.15}$$

Since this will be limited to cases where the gate voltage is significantly larger than the drain voltage, I neglect the gate-voltage dependence for this application, because

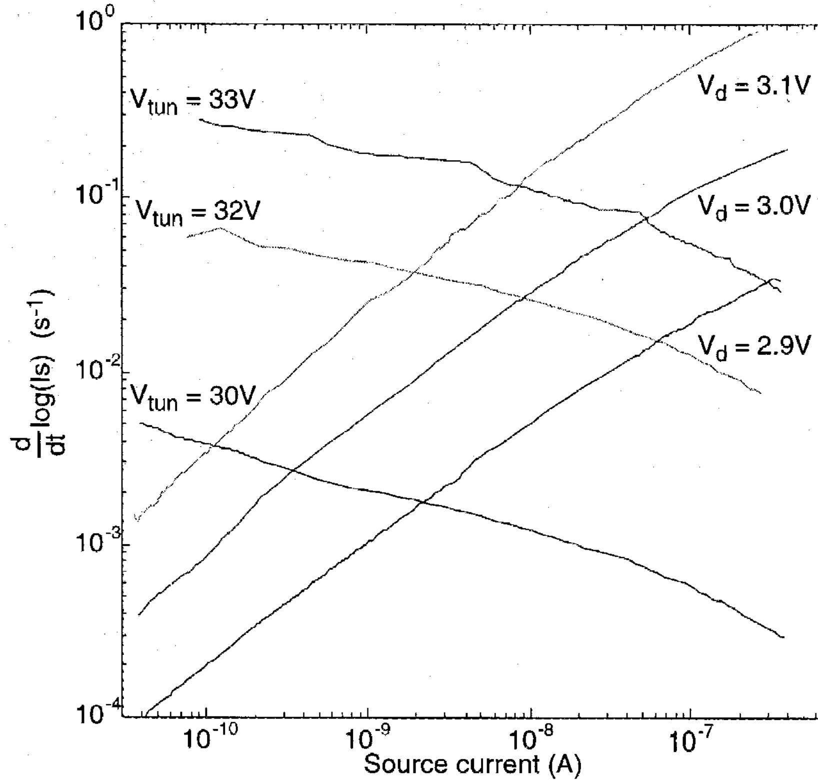


Figure 3.9: Plot of $\frac{d}{dt}\log(I_s)$ as a function of I_s for three different tunneling and three different drain voltages. Starting at an appropriately low (or high) source current (I_s), I measured this data by stepping to the desired drain or tunneling voltage, and letting the tunneling (or injection) current decrease (or increase) the synapse current. The drain voltage was 2V when tunneling, and the tunneling voltage was 23V when injecting. This measurement gives the floating-gate weight update rule as a function of I_s; plotting the data in this way shows the power-law dependence of channel current on the floating-gate current.

the injection current is only a weak function of the floating-gate voltage for a fixed source current (I_s) and Φ_{dc}. As seen in Fig. 3.8, f_2 is approximately linear over a 1V change in Φ_{dc}; therefore, around a quiescent level of Φ_{dc}, the injection current will e-fold for a Φ_{dc} increase of V_{inj}. Figure 3.8 also illustrates the nFET and pFET V_{inj} parameters and their range of validity.

When the gate voltage is the input modulating the channel current, the circuit model requires one more modification. A decreasing input signal will decrease the FET surface potential via capacitive coupling to the floating gate. Decreasing the

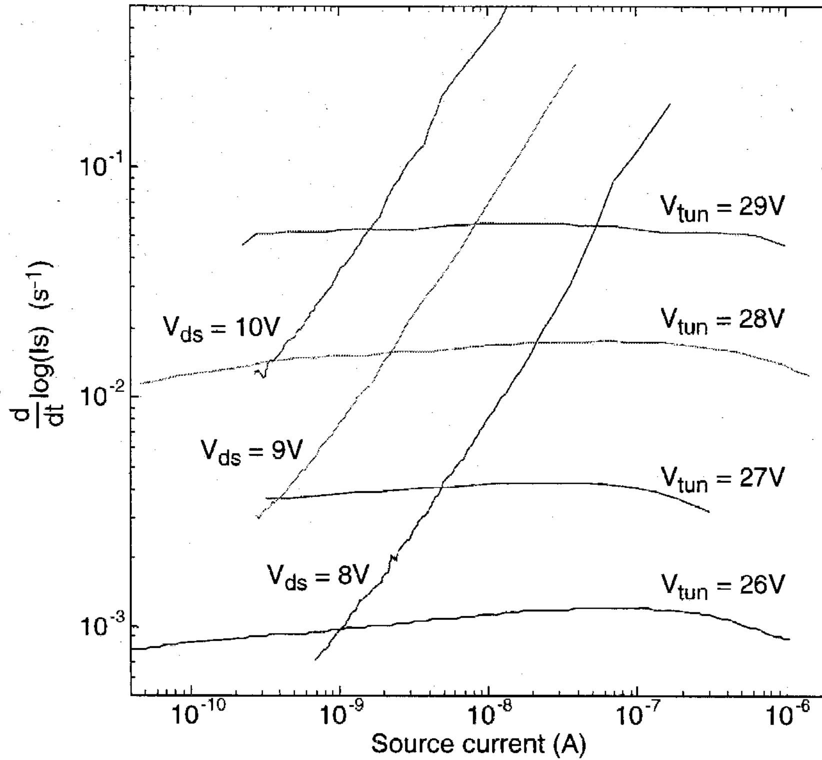


Figure 3.10: Plot of $\frac{d}{dt}\log(I_s)$ as a function of I_s for four different tunneling and three different drain voltages. Plotting the data in this way shows the power-law dependence of channel current on the floating-gate current. I measured this injection (tunneling) data as in Fig. 3.9 by starting the synapse at a low (or high) source current. Since this is a pFET synapse, the tunneling voltages are referenced to V_{dd}.

FET surface potential will increase the source current, thereby decreasing Φ_{dc} for a fixed output voltage, and lowering the injection efficiency. Using the linearized hot-electron–injection model, I model the injection current as the α power of the source current as follows:

$$\text{nFET:} \quad I_{inj} = I_{inj0}\left(\frac{I_s}{I_{s0}}\right)^{\alpha} \exp\left(\frac{\Delta V_d}{V_{inj}}\right),$$

$$\text{pFET:} \quad I_{inj} = I_{inj0}\left(\frac{I_s}{I_{s0}}\right)^{\alpha} \exp\left(-\frac{\Delta V_d}{V_{inj}}\right). \tag{3.16}$$

We can write (3.16) in terms of the gate voltage as

$$\begin{aligned} \text{nFET:} \quad & I_{inj} = I_{inj0}\text{exp}\left(\frac{\alpha\kappa\Delta V_{fg}}{U_T} + \frac{\Delta V_d}{V_{inj}}\right), \\ \text{pFET:} \quad & I_{inj} = I_{inj0}\text{exp}\left(-\frac{\alpha\kappa\Delta V_{fg}}{U_T} - \frac{\Delta V_d}{V_{inj}}\right), \end{aligned} \tag{3.17}$$

where I_{inj0} is the quiescent tunneling current, and α is $1 - \frac{U_T}{V_{inj}}$. A typical nFET value of α is 0.70, and a typical pFET value of α is 0.90; both values are consistent with typical values of V_{inj}. Since hot electron injection adds electrons to the floating gate, the current into the floating gate is negative, which results in the dynamical equations

$$\begin{aligned} \text{nFET:} \quad & \frac{U_T C_T}{\kappa_n I_{inj0}} \frac{dW}{dt} = (W)^{1+\alpha} \exp\left(\frac{\Delta V_d}{V_{inj}}\right), \\ \text{pFET:} \quad & -\frac{U_T C_T}{\kappa_p I_{inj0}} \frac{dW}{dt} = (W)^{1+\alpha} \exp\left(-\frac{\Delta V_d}{V_{inj}}\right). \end{aligned} \tag{3.18}$$

Figures 3.9 and 3.10 show the plot of $\frac{d}{dt}\log(I_s)$ for hot-electron injection as a function of I_s for different drain voltages. The straight line behavior validates the model in (3.18). The tunneling voltage was 20V during this hot-electron injection experiment. From (3.5), I plot the derivative of $\log(W)$ versus W to illustrate the dependence of source current on the floating-gate currents. Typical nFET values of α are in the range of $0.7 - 0.9$, while typical pFET values of α are in the range of $0.88 - 0.95$.

3.4 An Array of Single Transistor Synapses

Up to this point, we have only considered single independent synapses, but these synapses are intended to be coupled as an array of processors. The particular learning algorithm depends on the circuitry at the boundaries of the array; in particular the circuitry connected to each of the source, drain, and tunneling lines in a row. Chapter 4 partially addresses the possible learning rules in these synapses by considering the floating-gate dynamics in several fundamental floating-gate circuits. Chapter 4

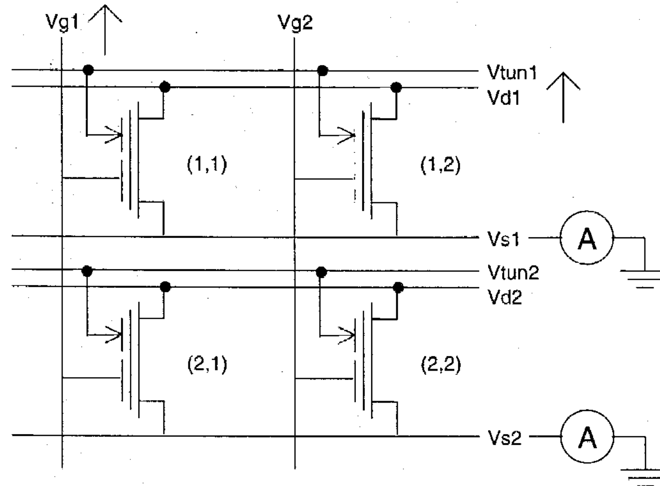


Figure 3.11: Circuit diagram of the single-transistor synapse array. Each transistor has a floating gate capacitively coupled to an input column line. A 2 x 2 section of the array allows us to characterize how modifying a single floating gate (such as synapse (1,1)) affects the neighboring floating gate values. The synapse currents are a measure of the synaptic weights, and are summed along each row by the source (V_s) or drain (V_d) lines into a typical soma circuit.

shows a variety of stabilizing, destabilizing, cooperative, and competitive floating-gate circuits that would be the basis of a learning network. Another issue is the network topology, that is, which terminals are connected along the rows and columns, and which terminals are the inputs, outputs, and error signals. I have built several topologies, each having advantages in a particular application. Figure 3.11 shows a commonly used topology of a 2x2 array of nFET synapses. For another example, a multilayer backpropagation network might want to connect the drain and tunneling terminals along a column and use these terminals as the input, and connect the source and gate terminals along a row. This topic is the subject of ongoing research and is beyond the scope of this thesis.

The remaining question is how the synapses interact when coupled into an array. The floating gate is localized to a particular device (confirmed by experimental measurements), as opposed to biological synapses where the efficacy modulators are locally diffused in space. Therefore, the device interactions are due entirely to the nonlinear dependence of the terminal voltages on the floating-gate current. I measure

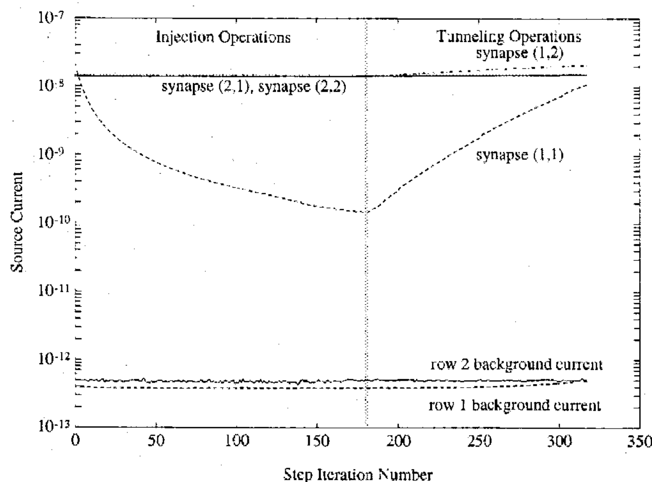


Figure 3.12: Output currents from a 2 x 2 section of the synapse array, showing 180 injection operations followed by 160 tunneling operations. Because our measurements from the 2 x 2 section come from a larger array, we also display the 'background' current from all other synapses on the row. This background current is several orders of magnitude smaller than the selected synapse current, and therefore negligible.

this effect by considering how selectively we can modify the charge on a particular floating gate without affecting the other floating gates. Figure 3.12 shows these results from the synapse array shown in Fig. 3.11. The experiment performed 180 injection operations followed by 160 tunneling operations. For the injection operations, the drain (V_{d1}) is pulsed from 2.0V up to 3.3V for 0.5s with V_{g1} at 8V and V_{g2} at 0V. For the tunneling operations, the tunneling line (V_{tun1}) is pulsed from 20V up to 33.5V with V_{g2} at 0V and V_{g1} at 8V.

Figure 3.13(a) shows measured data on the change in source current during tunneling as a function of source current for several values of tunneling voltage. The tunneling operation increases the synaptic weight. V_{g1} was held at 0V and V_{g2} was 8V while the tunneling line was pulsed for 0.5s from 20V to the voltage shown. The change in source current is approximately proportional to the 1 - $\frac{U_T}{\kappa_n V_x}$ power of the source current where $\frac{U_T}{\kappa_n V_x}$ is between 0.1 and 0.3 for the range of tunneling voltages shown. The effect of this tunneling procedure on synapse (2,1) and (2,2) are negligible.

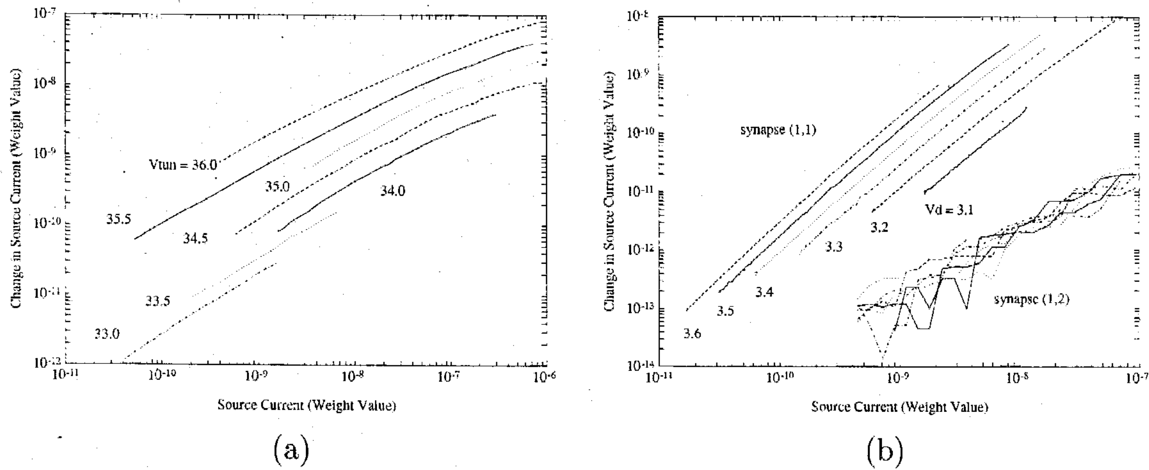


Figure 3.13: (a) Synapse (1,1) source current increment versus source current for several values of tunneling voltage. (b) Source current decrement during injection versus source current for several values of drain voltage.

Tunneling selectivity along a row in this array is entirely a function of how far apart the two floating gates are pushed by the gate inputs. To select a particular synapse, we want to bring that floating gate to as low a voltage as possible, while at the same time bringing all the other floating gates to as high a voltage as possible. The selectivity ratio of synapses on the same row is typically between 3-7 for our devices. The tunneling selectivity can be increased by increasing the input voltage steps or by increasing the gate coupling to the floating gate. A low or high gate coupling for these synapses is entirely application dependent; the circuits in Chapter 5 decrease intentionally this ratio to increase the circuit's linear and dynamic range.

Figure 3.13(b) shows measured data on the change in source current during injection versus source current for several values of drain voltage. The injection operation decreases the synaptic weight. V_{g2} was held at 0V, and V_{g1} was at 8V during the 0.5s injecting pulses. The change in source current is approximately proportional to the source current to the $1 + \alpha$ power, where α is between 0.7 and 0.85 for the range of drain voltages shown. The change in source current in synapse (1,2) is much less than the corresponding change in synapse (1,1) and is nearly independent of drain voltage. The injection operations resulted in negligible changes in source current for

synapses (2,1) and (2,2). Ideally, a synaptic array is programmed (or initialized) by injection after the synaptic array is set to a desired baseline by tunneling, because injection selectivity is much greater than tunneling selectivity,

In conclusion, we present an approximate model of our array of these single transistor synapses. The learning increment of the synapse at position (i, j) can be modeled as for the nFET synapse as

$$I_{s_{i,j}} = I_{so} W_{i,j} \exp\left(\frac{\kappa_n}{U_T}\frac{C_1}{C_T}\Delta\hat{V}_{in_i} + \frac{\kappa_n}{U_T}\frac{C_2}{C_T}\Delta\hat{V}_{d_j}\right),$$

$$\frac{U_T C_T}{\kappa_n W_{i,j}}\frac{dW_{i,j}}{dt} = C_1\frac{d\bar{V}_{in_i}}{dt} + C_2\frac{d\bar{V}_{d_j}}{dt} + (W_{i,j})^{-\frac{U_T}{\kappa_n V_x}} - (W_{i,j})^{\alpha}\exp\left(\frac{\Delta V_{d_j}}{V_{inj}}\right), \quad (3.19)$$

and for the pFET synapse as

$$I_{s_{i,j}} = I_{so} W_{i,j} \exp\left(-\frac{\kappa_p}{U_T}\frac{C_1}{C_T}\Delta\hat{V}_{in} - \frac{\kappa_p}{U_T}\frac{C_2}{C_T}\Delta\hat{V}_{d}\right),$$

$$\frac{U_T C_T}{\kappa_p W_{i,j}}\frac{dW_{i,j}}{dt} = -C_1\frac{d\bar{V}_{in_i}}{dt} - C_2\frac{d\bar{V}_{d_j}}{dt} - (W_{i,j})^{\frac{U_T}{\kappa_p V_x}} + (W_{i,j})^{\alpha}\exp\left(-\frac{\Delta V_{d_j}}{V_{inj}}\right). \quad (3.20)$$

Typical values for the parameters have been defined earlier in this chapter.

Chapter 4 Continuous-Time Feedback in Floating-Gate MOS Circuits

Although Chapter 3 presented the electron-tunneling, hot-electron–injection, and multiplicative behavior of the single-transistor synapses, and derived effective learning rules of single-transistor synapses, it did not consider any specific ways that these synapses could be used in a network. This chapter considers the behaviors that emerge when single-transistor synapses are coupled together to form various continuous-time learning networks. My purpose is to understand the dynamics of the learning mechanisms naturally available in floating-gate MOS circuits (FGMOS). Chapter 5 presents the autozeroing floating-gate amplifier (AFGA), which is the first circuit application of a single-transistor synapse with continuous oxide currents.

This chapter describes the various possible negative- and positive-feedback mechanisms in continuous-time FGMOS circuits. The usefulness of negative or positive feedback depends on the application; Hebbian learning [83], for example, is a case of destabilizing positive feedback. Section 4.1 presents a new type of pFET synapse, in which the gate currents provide stabilizing feedback to the floating gate and to the drain. Section 4.2 presents the range of possible stabilizing and destabilizing feedback configurations in circuits comprising one floating-gate synapse; we include data from nFET, pFET, and source-degenerated pFET synapses. Section 4.3 presents the class of stable two-synapse circuits by showing examples of competitive and cooperative behavior between synapses; these properties extend directly to networks of many synapses. I present my conclusions in Section 4.4.

The small-signal models of the single-transistor synapse elements often shed considerable light on the dynamics of simple FGMOS circuits. Figure 4.1 shows small-signal models of the nFET and pFET synapses, where all the small-signal parameters are positively valued. Table 4.1 shows the relationships of the small-signal parame-

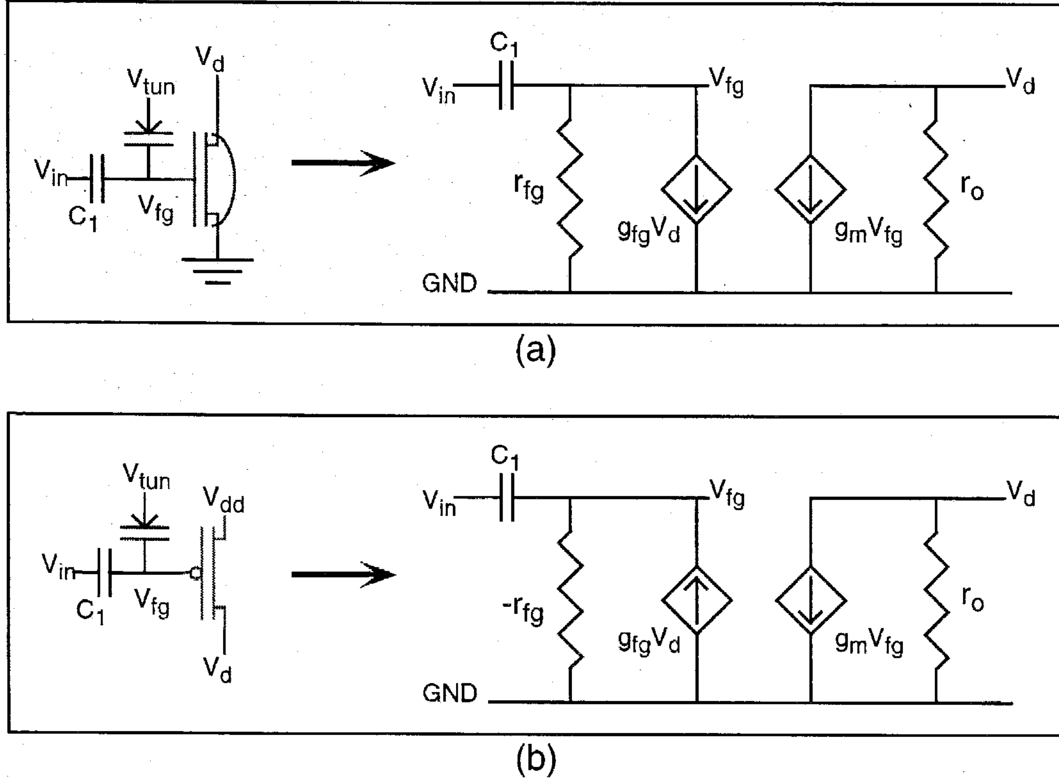


Figure 4.1: (a) Circuit diagram and small-signal model of the nFET single-transistor synapse with its source connected to ground. The small-signal model assumes a constant tunneling current and that the parameters are positive. (b) Circuit diagram and small-signal model of the pFET single-transistor synapse with its source connected to V_{dd}. The small-signal model assumes a constant tunneling current and that the parameters are positive. The small-signal resistance from floating-gate to ground is negative due to the hot-electron injection currents.

ters to the device parameters; I define $x||y$ as $\frac{xy}{x+y}$. I use the conventional definitions [81] of small-signal transconductance, g_m, and output resistance, r_o, to compute the values in Table 4.1 from (3.1). I define this transistor's maximum voltage gain, A_v, as the product $g_m r_o$. I used (3.11) and (3.16) to calculate the values of g_{fg} and r_{fg} in Table 4.1. I define g_{fg} as the change in the gate current in response to a change in drain voltage. Because only the injection current depends on the drain voltage, $g_{fg} = \frac{I_{tun0}}{V_{inj}}$. I define r_{fg} as the magnitude of the change in gate current for a change in gate voltage. The resistance from the floating gate to ground is negative for a pFET. A typical nFET value for the product $g_{fg}r_{fg}$—the gain from V_d to V_{fg}—is 1; a typical

Table 4.1: Relationships of small-signal parameters

Parameter	nFET	pFET	s-d pFET
$g_m = \lvert \frac{\partial I_s}{\partial V_{fg}} \rvert$	$\frac{\kappa_n I_{so}}{U_T}$	$\frac{\kappa_p I_{so}}{U_T}$	$\frac{\kappa_p \kappa_x I_{so}}{U_T}$
$r_o = \lvert \frac{\partial I_s}{\partial V_d} \rvert$	$\frac{V_o}{I_{so}}$	$\frac{V_o}{I_{so}}$	$\frac{V_o}{\kappa_x I_{so}}$
$A_v = g_m r_o$	$\frac{\kappa_n V_o}{U_T}$	$\frac{\kappa_p V_o}{U_T}$	$\frac{\kappa_p V_o}{U_T}$
$g_{fg} = \lvert \frac{\partial I_{inj}}{\partial V_d} \rvert$	$\frac{I_{tun0}}{V_{inj}}$	$\frac{I_{tun0}}{V_{inj}}$	$\frac{I_{tun0}}{V_{inj}}$
$r_{fg} = \lvert \frac{\partial (I_{tun}+I_{inj})}{\partial V_{fg}} \rvert$	$\frac{U_T/\kappa_p \parallel V_x}{I_{tun0}}$	$\frac{U_T/\kappa_p \parallel -V_x}{I_{tun0}}$	$\frac{U_T/\sigma \parallel V_x}{I_{tun0}}$
$A_{fg} = g_{fg} r_{fg}$	$\frac{U_T/\kappa_p \parallel V_x}{V_{inj}}$	$\frac{U_T/\kappa_p \parallel -V_x}{V_{inj}}$	$\frac{U_T/\sigma \parallel V_x}{V_{inj}}$

pFET value is 0.3.

4.1 The Source-Degenerated pFET Synapse

This section presents the source-degenerated pFET synapse, which I show in Fig. 4.2. I model the κ_x-diode element as

$$I = I_x \exp\left(-\frac{\kappa_x \Delta V}{U_T}\right), \tag{4.1}$$

where ΔV represents the change in V from the bias level, I_x is the quiescent current through the device, and κ_x is a device parameter. By equating the current through the pFET and the κ_x-diode element, and substituting this expression into (3.1), we obtain the source-degenerated pFET model equation for $\kappa_x \gg 1$, which is our region of interest:

$$I_s = I_{so} \exp\left(-\kappa_x \left(\frac{\kappa_p V_{fg}}{U_T} + \frac{V_d}{V_o}\right)\right). \tag{4.2}$$

Figure 4.3 shows the channel current through this synapse as a function of gate voltage for three sizes of κ_x; reducng κ_x decreases the change in channel current for a fixed gate-voltage swing. The κ_x element is effectively cascoded by the pFET; therefore, from (4.2), the maximum voltage gain of this transistor is $\frac{\kappa V_o}{U_T}$, which is identical to that of a non-degenerated subthreshold pFET transistor.

As in (3.3), I can separate the terminal voltages into fast and slow variables. The

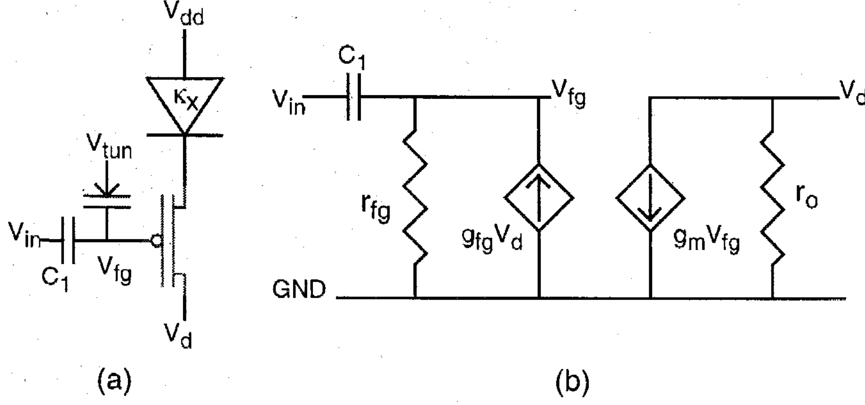


Figure 4.2: The source-degenerated pFET single-transistor synapse. (a) Circuit of the source-degenerated pFET synapse. (b) The small-signal of the source-degenerated pFET synapse, where I have defined all of the small signal quantities to be positive. From this circuit, we can see that this pFET's floating-gate currents provide stabilizing feedback to the floating-gate and drain voltages.

fast-variable definitions are identical to the pFET fast-variable definitions in (3.4) and (3.7), except that κ_p is replaced by $\kappa_p \kappa_x$; the slow-variable model is

$$-\frac{C_T U_T}{\kappa_x \kappa_p W} \frac{dW}{dt} = C_1 \frac{dV_{in}}{dt} + C_2 \frac{d\bar{V}_{out}}{dt} + I_{tun} - I_{inj}, \tag{4.3}$$

where I_{tun} and I_{inj} are the electron-tunneling and hot-electron–injection currents of the pFET.

Figures 4.4, 4.5, and 4.6 show $\frac{d}{dt}\log(I_s)$ versus the source current, I_s, for various tunneling and hot-electron injection currents. Figure 4.2 shows the equivalent small-signal circuit, and Table 4.1 presents the small-signal parameters for this source-degenerated pFET. We can express the change in the tunneling current in terms of the source current, by substituting (4.2) into (3.9):

$$I_{tun} = I_{tun0} \left(\frac{I_s}{I_{so}} \right)^{\frac{U_T}{\kappa_p \kappa_x V_x}} . \tag{4.4}$$

From Fig. 4.5, the tunneling current is nearly a constant power of I_s, and the power increases with decreasing κ_x.

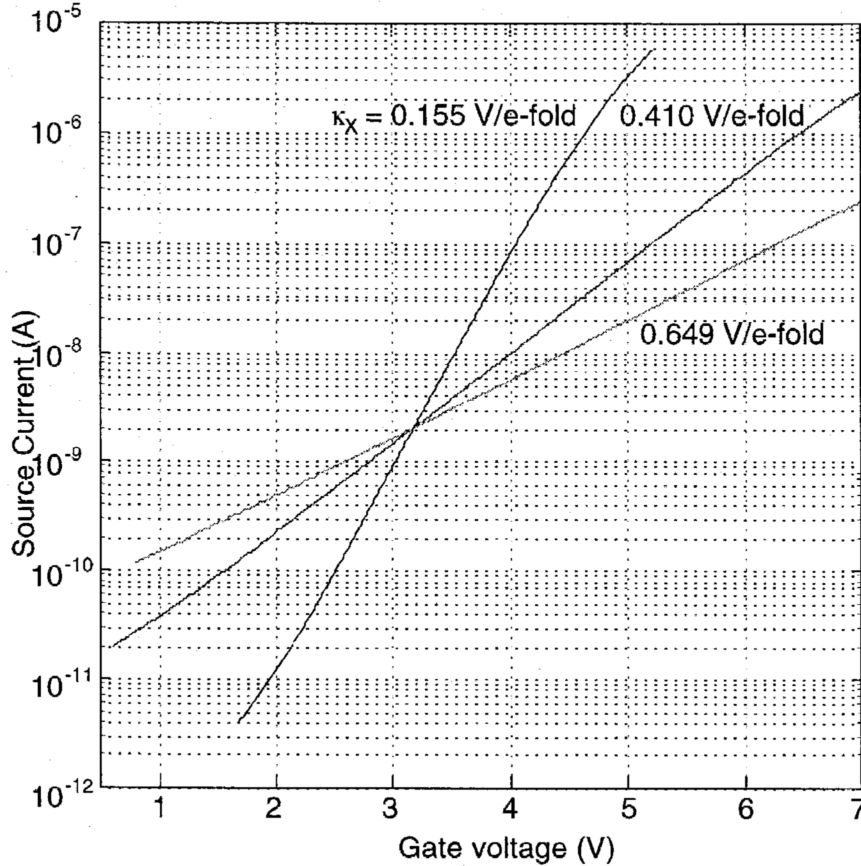


Figure 4.3: Plot of source-degenerated pFET drain current versus gate voltage for three values of κ_x. Decreasing κ_x decreases the transistor's effective gate coupling to the surface potential (due to the decreasing exponential slope).

We can derive the hot-electron injection model for this circuit by substituting the source-degenerated pFET model (4.2) into the pFET hot-electron injection model (3.16):

$$I_{inj} = I_{inj0} \exp\left(\frac{\sigma \Delta V_g}{U_T}\right) \exp\left(-\frac{\Delta V_d}{V_{inj}}\right), \tag{4.5}$$

where $\sigma = \kappa_p \left(\frac{U_T}{V_{inj}} - \kappa_x \alpha\right)$, for small κ_x. The value of κ_x strongly affects the source-current dependence of the injection current. For small κ_x, an increase in the source current requires a large increase in the source voltage, which decreases the drain-to-source voltage, and that, in turn decreases the injection current. With no feedback,

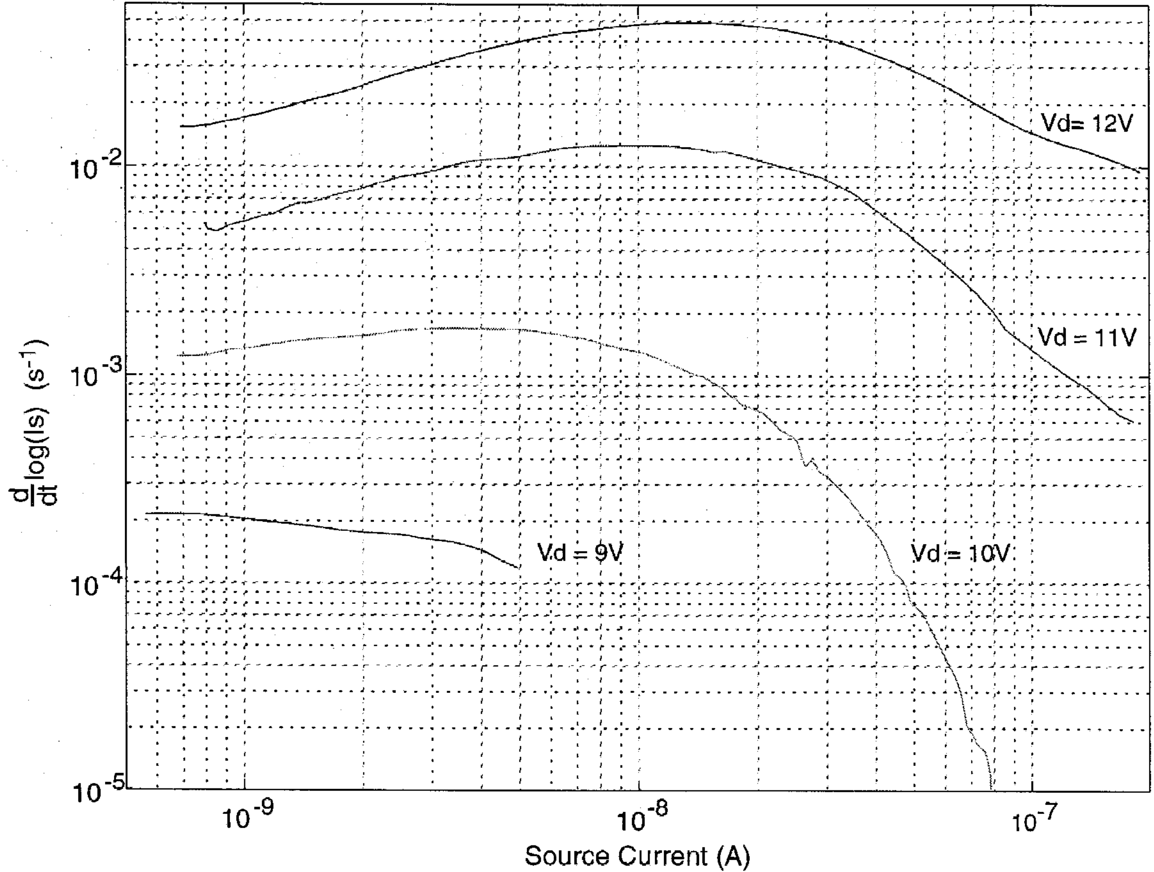


Figure 4.4: Plot of $\frac{d\log(I_s)}{dt}$ versus source current (I_s) for four different values of drain voltage in the source-degenerated pFET. The transistor has stabilizing behavior for some regions of this graph; these portions are reflected in the small signal model. The synapse goes from unstable feedback to stable feedback due to the change in the bias Φ_{dc} of the synapse; larger Φ_{dc} requires a lower κ_x for stabilizing behavior.

σ is negative; the value of κ_x at which σ changes from negative to positive is given by

$$\kappa_x = \frac{U_T}{\alpha V_{inj}}. \tag{4.6}$$

Therefore, if V_{inj} increases, κ_x must decrease by the same amount to preserve stability. By writing the injection current as a function of the source current, we get

$$I_{inj} = I_{inj0} \left(\frac{I_s}{I_{so}}\right)^{-\frac{\sigma}{\kappa_p \kappa_x}} \exp\left(-\frac{\Delta V_d}{V_{inj}}\right). \tag{4.7}$$

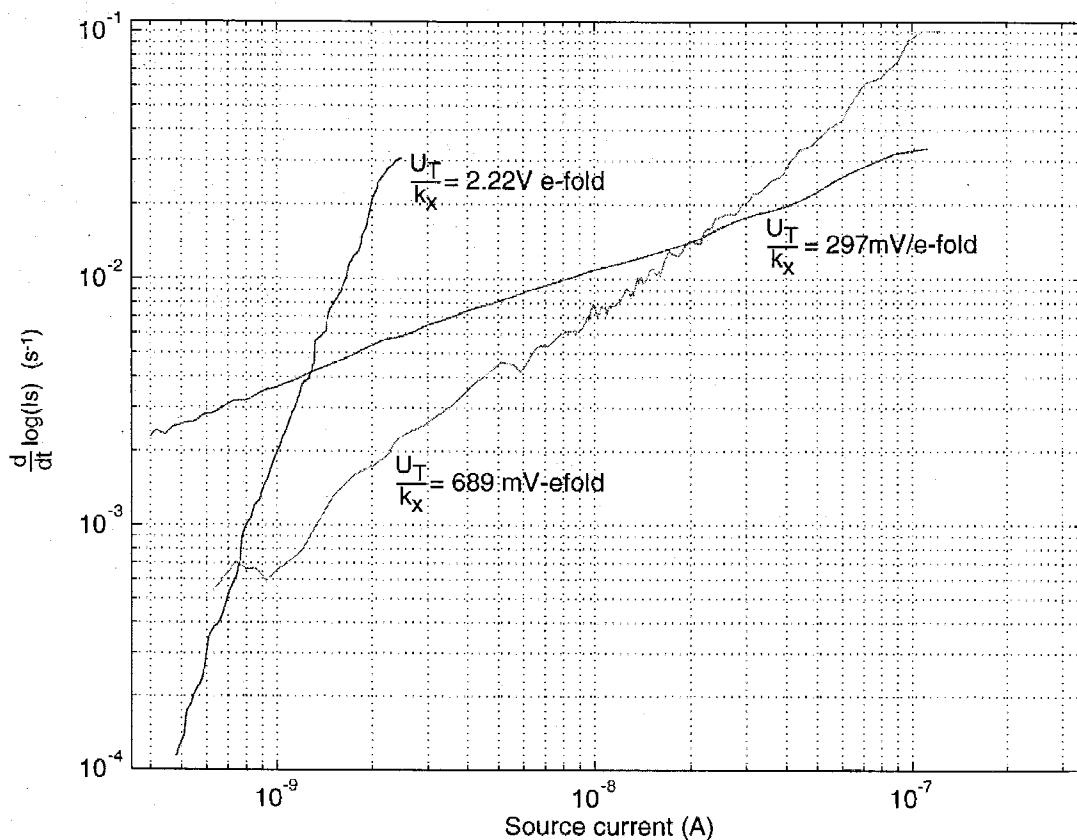


Figure 4.5: The effect of κ_x on tunneling current in the source-degenerated pFET. Plot of $\frac{d}{dt}\log(I_s)$, which is proportional to the tunneling current, as a function of I_s for three different values of κ_x. The curves are nearly straight lines, and the slope of the curves increases for a decreasing κ_x.

Figure 4.4 plots $\frac{d}{dt}\log(I_s)$, as a function of I_s for three values of drain voltage; Figure 4.6 plots $\frac{d}{dt}\log(I_s)$, as a function of I_s for three values of κ_x. The injection current is not a constant power of I_s, because V_{inj} increases when Φ_{dc} increases. An nFET synapse with a κ_x source degeneration would show similar changes in stability.

To obtain the data in Figs. 4.4 and 4.5, I set κ_x by an integrated pnp transistor and an op-amp feedback circuit; this approach gave me the freedom to investigate different values of κ_x. In an integrated synapse, I use a short-channel pFET with significant drain-induced barrier lowering (DIBL) to implement the κ_x diode-element. A transistor that strongly exhibits DIBL shows an exponential change in current for a linear change in drain voltage, as I show in Fig. 4.7. This approach may suffer from significant mismatch, because to make a FET with strong DIBL effect, one

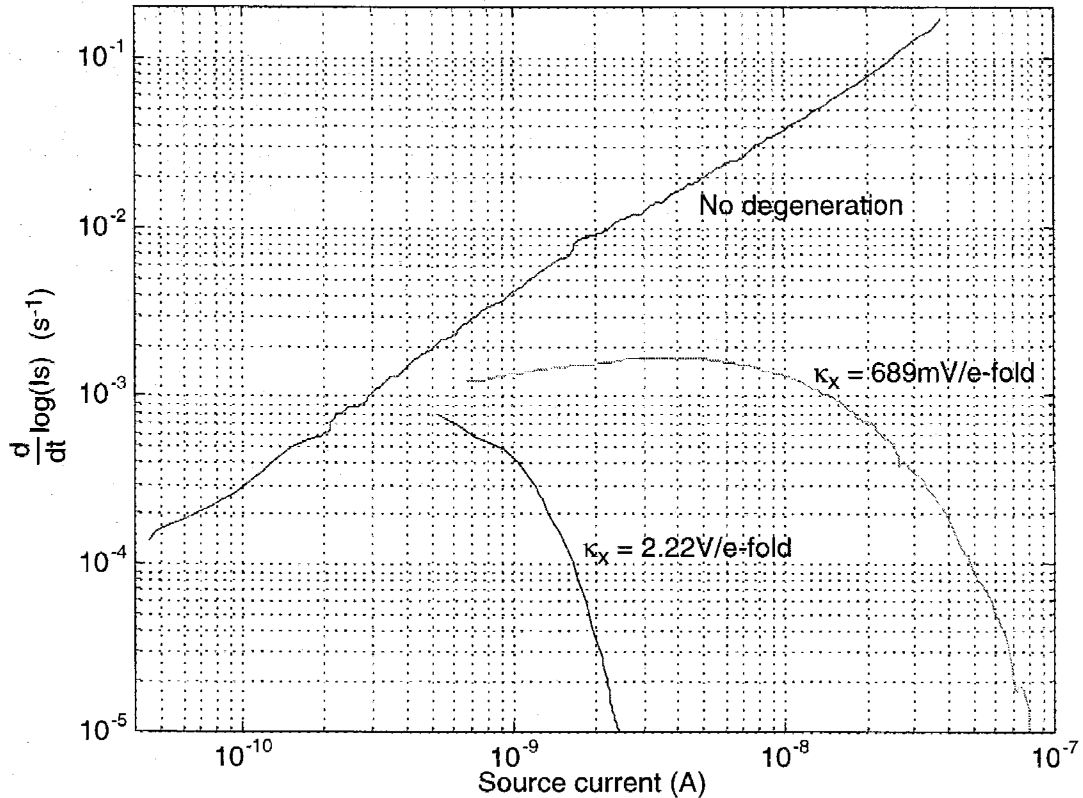


Figure 4.6: The effect of κ_x on hot-electron–injection current in the source-degenerated pFET. Plot of $\frac{d}{dt}\log(I_s)$, which is proportional to the injection current, as a function of I_s for three different values of κ_x, including $\kappa_x = 0$.

must violate the process design rules. An alternative approach is to set κ_x using a floating-gate pFET that has a floating-gate–to drain capacitance much smaller than the total capacitance of the floating gate; however, this scheme requires a method of initializing the floating-gate charge.

4.2 Stability of Single-Synapse Circuits with Floating-Gate Feedback

This section considers simple one-synapse circuits, to illustrate the basic floating-gate feedback mechanisms. I consider synapse configurations that couple through the drain terminal; similar feedback mechanisms occur with other configurations that employ

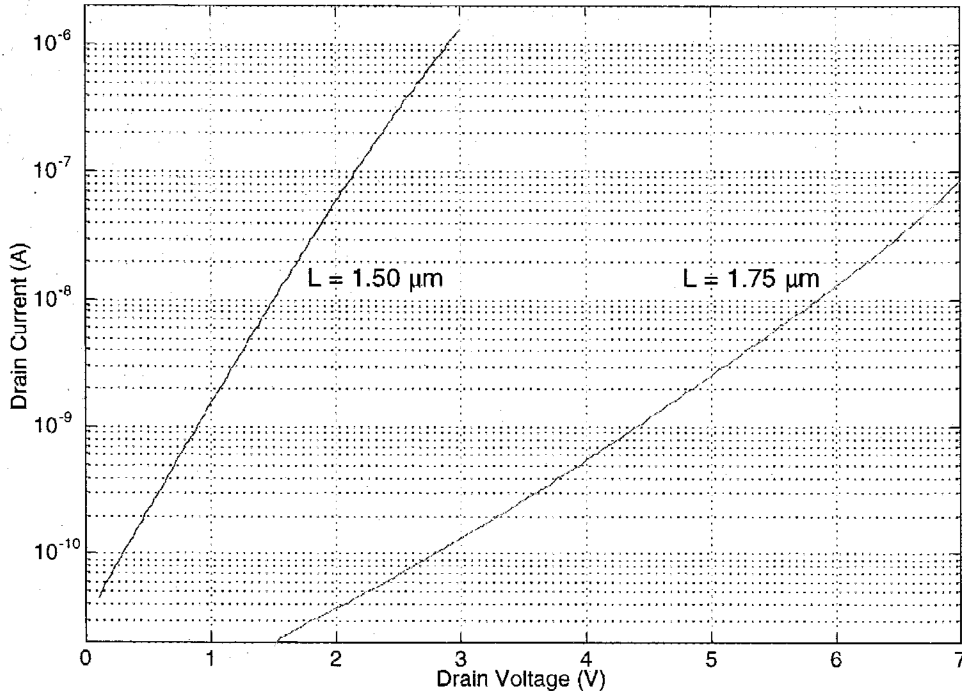


Figure 4.7: Source current versus drain voltage for two short-channel pFETs in a 2μm process with the gate and source voltages at V_{dd}. The drain voltage is measured relative to V_{dd}. The DIBL results in an exponentially increasing current for subthreshold biases; for the 1.5μm pFET, the current increases an e-fold for a 276.1mV change in the drain voltage, and for the 1.75μm pFET, the current increases an e-fold for a 627.3mV change in the drain voltage.

various combinations of the drain, source, and tunneling terminals. I only consider sufficiently long timescales such that the current through the synapse is identical to the current through the coupling FET; that is, I ignore capacitive currents at the drain and source nodes. I also assume that the synapses have nonnegligible tunneling and injection currents.

Figures 4.8 and 4.11 show the pFET and nFET circuits that comprise a single synapse with the the drains connected to current sources. Figures 4.9 and 4.12 show the pFET and nFET circuits that comprise a single synapse with the drains connected to cascode transistors. For all four circuits, equilibrium is established when the tunneling current is balanced by the injection current. Figure 4.10 shows how to simplify the small-signal models for the pFET circuits; one can similarly simplify

the nFET and source-degenerated pFET circuits. Using a fixed channel current is equivalent to open-circuiting the drain terminal in the synapse's small-signal model; therefore, the effective conductance from floating gate to ground is

$$n\text{FET} : r_{fg} - (g_m r_o)\, g_{fg},$$

$$p\text{FET} : (g_m r_o)\, g_{fg} - r_{fg}. \tag{4.8}$$

Since this effective conductance is positive for pFET synapses, the configuration in Fig. 4.8 is stable. On the other hand, since this effective conductance is negative for nFET synapses, the configuration in Fig. 4.11 is unstable. When the drain voltage is fixed, the sign of r_{fg} determines the stability of this circuit; the configuration in Fig. 4.9 is unstable due to the negative r_{fg} for pFET synapses, whereas the configuration in Fig. 4.12 is stable due to the positive r_{fg} for nFET synapses. For both the pFET and nFET synapses, a particular load resistance, R_l, connected to the drain, where $R_l g_m (g_{fg} r_{fg}) = 1$, results in zero effective conductance between the floating gate and ground and is the boundary between the stable and unstable regimes. Because this formulation is valid for positive and negative resistances connected to the drain, negative drain resistance is stablizing for nFETs and destabilizing for pFETs. For sufficiently small κ_x, the source-degenerated pFET synapse has a positive conductance from floating gate to ground for all positive resistances connected to the drain terminal, as seen in Table 4.1 and Fig. 4.4. The following three subsections consider the detailed behavior of the pFET, nFET, and source-degenerated pFET circuits.

4.2.1 *p*FET Floating-Gate Circuits

Figure 4.8 shows an autozeroing floating-gate amplifier (AFGA). Chapter 5 explains how the AFGA behaves as a bandpass amplifier; here, I summarize the AFGA's adaptation properties. With capacitive feedback, the input signal is amplified by a closed-loop gain approximately equal to $-\frac{C_1}{C_2}$, where C_2 is the capacitance between the floating gate and drain. The complementary tunneling and hot-electron injection

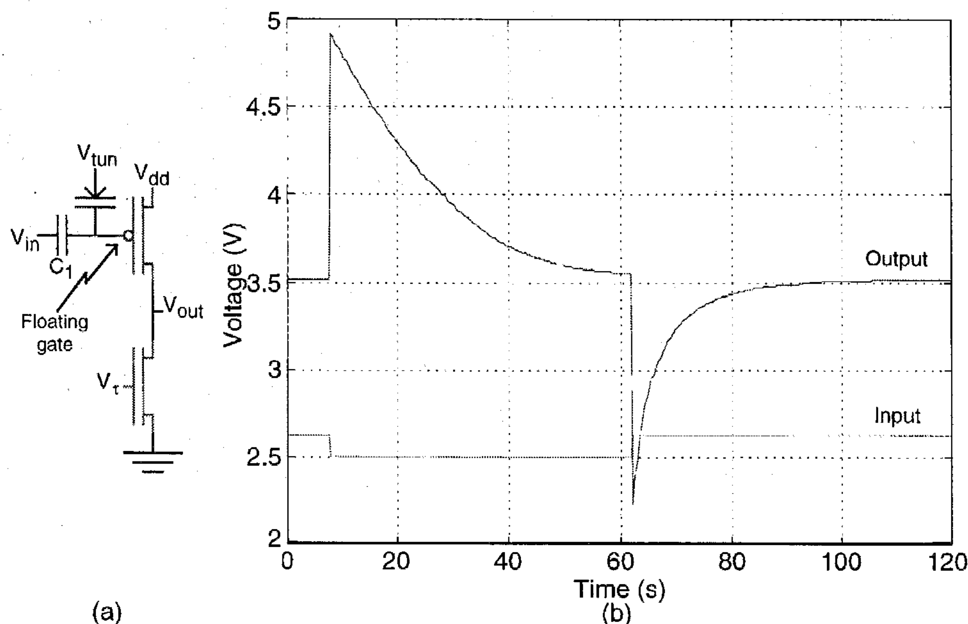


Figure 4.8: (a) The pFET voltage-adapting circuit configuration; this circuit is the autozeroing floating-gate amplifier (AFGA). (b) Response of the AFGA to an upgoing and a downgoing step input. The adaptation in response to an upward step results from electron tunneling; the adaptation in response to a downward step results from pFET hot-electron injection. This amplifier has a gain of 11.2, and I_{tun0} is 50fA.

processes adjust the floating-gate charge such that the amplifier's output voltage returns to a steady-state value on a slow time scale. If the output voltage is below its equilibrium value, then the injection current exceeds the tunneling current, decreasing the charge on the floating gate, and, in turn, raising the output voltage back toward its equilibrium value. If the output voltage is above its equilibrium value, then the tunneling current exceeds the injection current, increasing the charge on the floating gate, and, in turn, lowering the output voltage back toward its equilibrium value.

Because the amplifier has a large open-loop gain, if we are to keep the output voltage between the supply rails, the floating-gate voltage must be confined to a swing of only a few millivolts. A nearly constant floating-gate voltage implies a constant tunneling current ($I_{tun} = I_{tun0}$), and implies that the source current ($I_s = I_{so}$) is

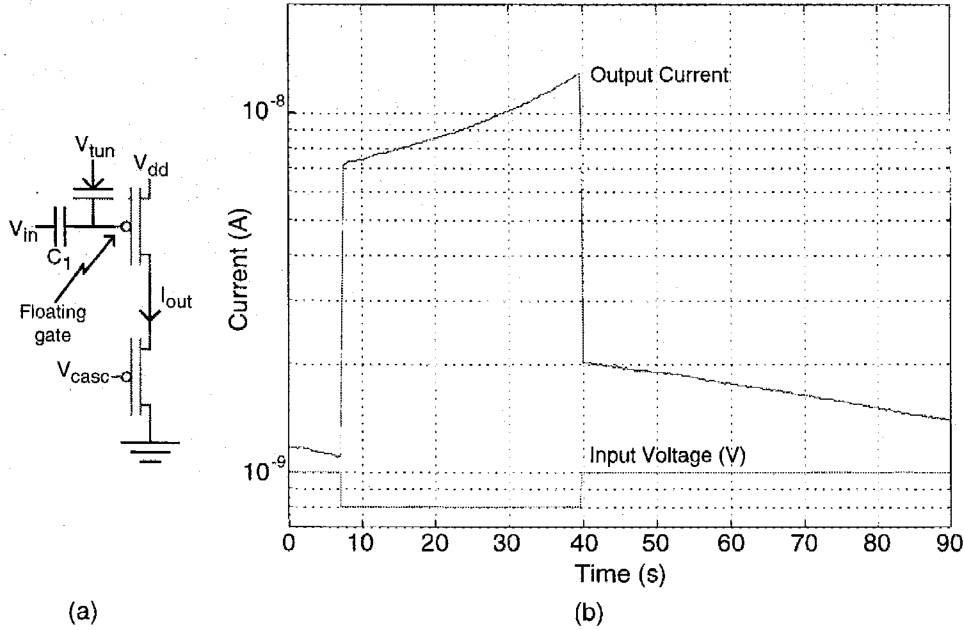


Figure 4.9: (a) The pFET current-adapting circuit configuration. (b) Response of the current-adapting pFET synapse to an upgoing and a downgoing step input. This circuit configuration is unstable.

equal to the nFET source current. Therefore, $W = 1$, and (3.7) simplifies to

$$C_2 \frac{dV_{out}}{dt} = -C_1 \frac{dV_{in}}{dt} + I_{tun0} \left(\exp\left(-\frac{\Delta V_{out}}{V_{inj}} \right) - 1 \right). \tag{4.9}$$

As I show in Chapter 5, (4.9) is a linear, first-order differential equation in X, where I define $X = \exp\left(\frac{\Delta V_{out}}{V_{inj}}\right)$. The trajectories of (4.9) converge to the steady state at $\Delta V_{out} = 0$. Figure 4.8 shows the circuit response to an upgoing and a downgoing input step; between the step changes in V_{in}, the output voltage converges back toward equilibrium, as expected from stabilizing feedback.

Figure 4.9 shows the pFET synapse circuit with its drain connected to a cascode transistor; the cascode configuration nearly fixes the drain voltage. If the output current is above its equilibrium value ($W = 1$), then the injection current exceeds the tunneling current, decreasing the floating-gate voltage, and further increasing the output current. If the output current is below its equilibrium value, then the

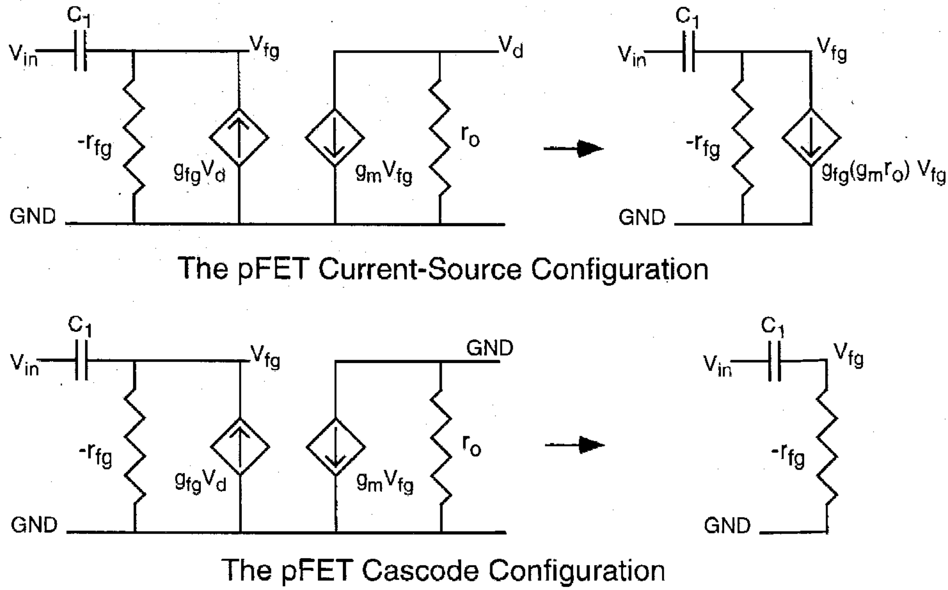


Figure 4.10: Simplification of small-signal models for the *p*FET single-transistor circuits; the *n*FET and source-degenerated *p*FET synapse circuits are simplified similarly. The top figure shows the simplified small-signal model for the drain connected to a cascode transistor. The bottom figure shows the simplified small-signal model for the drain connected to a current source. This circuit is unstable because of the negative resistance from floating-gate to ground.

tunneling current exceeds the injection current, increasing the floating-gate voltage, and further decreasing the output current. Because drain voltage is fixed, we can simplify (3.7) to

$$\frac{C_T U_T}{\kappa_p I_{tun0}} \frac{dW}{dt} = W \frac{C_1}{I_{tun0}} \frac{dV_{in}}{dt} - W^{1-\frac{U_T}{\kappa_p V_x}} + W^{1+\alpha}. \tag{4.10}$$

The trajectories of this differential equation converge toward $W = 0$ and diverge away from $W = 1$; this behavior is consistent with destabilizing positive feedback. Figure 4.9 shows the circuit response to input steps near the steady-state current; between the step changes in V_{in}, the output current diverges away from the equilibrium current.

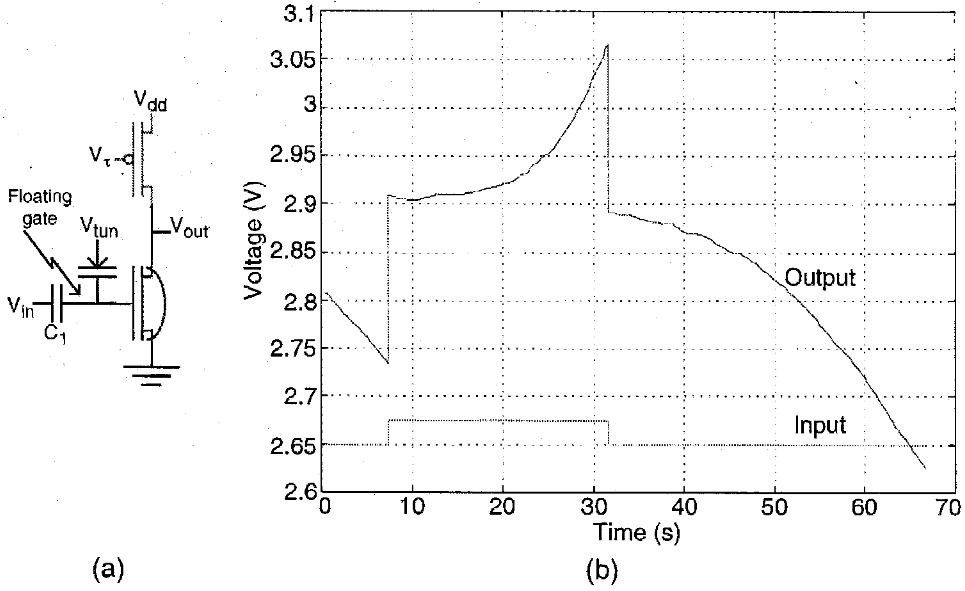


Figure 4.11: (a) The nFET voltage-adapting circuit configuration. (b) Response of the voltage-adapting nFET synapse to an upgoing and a downgoing step input. This circuit configuration is unstable.

4.2.2 nFET Floating-Gate Circuits

Figure 4.11 shows the nFET synapse with its drain connected to a current source. As with the AFGA, the source and tunneling currents are nearly constant, and W is fixed at 1. If the output voltage is above its equilibrium value ($\Delta V_{out} = 0$), then the injection current exceeds the tunneling current, decreasing the charge on the floating-gate, and further increasing the output voltage. If the output voltage is below its equilibrium value, then the tunneling current exceeds the injection current, increasing the charge on the floating-gate, and further decreasing the output voltage. The governing equation is

$$C_2 \frac{dV_{out}}{dt} = -C_1 \frac{dV_{in}}{dt} + I_{tun0} \left(\exp\left(\frac{\Delta V_{out}}{V_{inj}} \right) - 1 \right), \tag{4.11}$$

where the trajectories of this differential equation diverge away from the steady state at $\Delta V_{out} = 0$; this behavior is consistent with destabilizing positive feedback. Figure

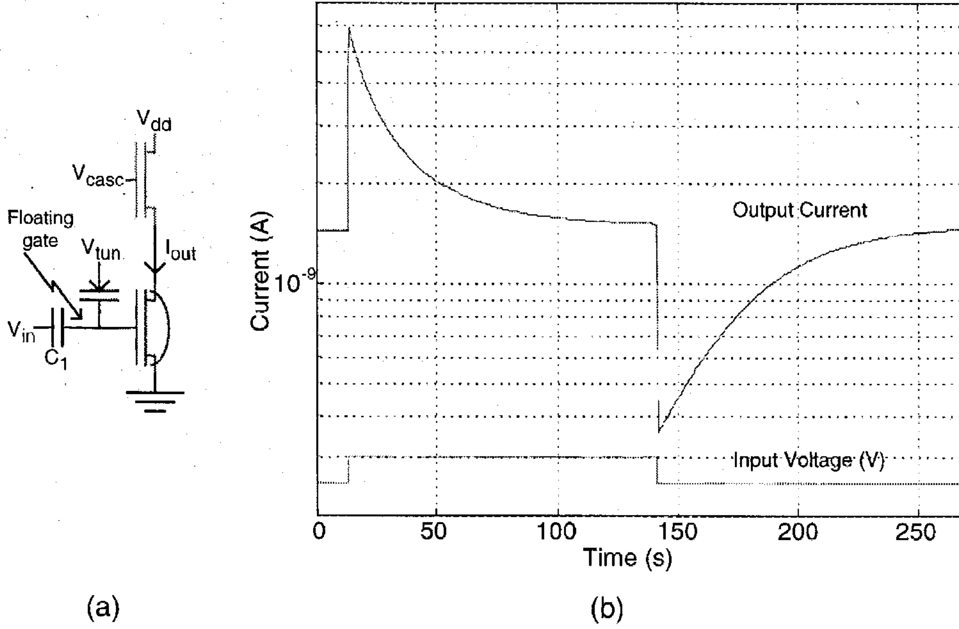


Figure 4.12: (a) The nFET current-adapting circuit configuration. (b) Response of the current-adapting nFET synapse to an upgoing and a downgoing step input. This circuit configuration is stable.

4.11 shows the circuit response to input steps near the steady-state voltage; between the step changes in V_{in}, the output voltage moves away from the equilibrium voltage.

Figure 4.12 shows the nFET synapse circuit configuration with a fixed drain voltage. If the output current is above its equilibrium value ($W = 1$), then the injection current exceeds the tunneling current, decreasing the floating-gate voltage, and in turn decreasing the output current. If the output current is below its equilibrium value, then the tunneling current exceeds the injection current, increasing the floating-gate voltage, and in turn increasing the output current. Because the drain voltage of the nFET synapse is fixed, I obtain from (3.7) that

$$\frac{C_T U_T}{\kappa_n I_{tun0}} \frac{dW}{dt} = W \frac{C_1}{I_{tun0}} \frac{dV_{in}}{dt} + W^{1-\frac{U_T}{\kappa_n V_x}} - W^{1+\alpha}; \tag{4.12}$$

the trajectories of this equation converge toward $W = 1$ and diverge away from $W = 0$; this behavior is consistent with stabilizing negative feedback. Figure 4.12 shows the circuit response to input steps near the steady-state voltage; between the

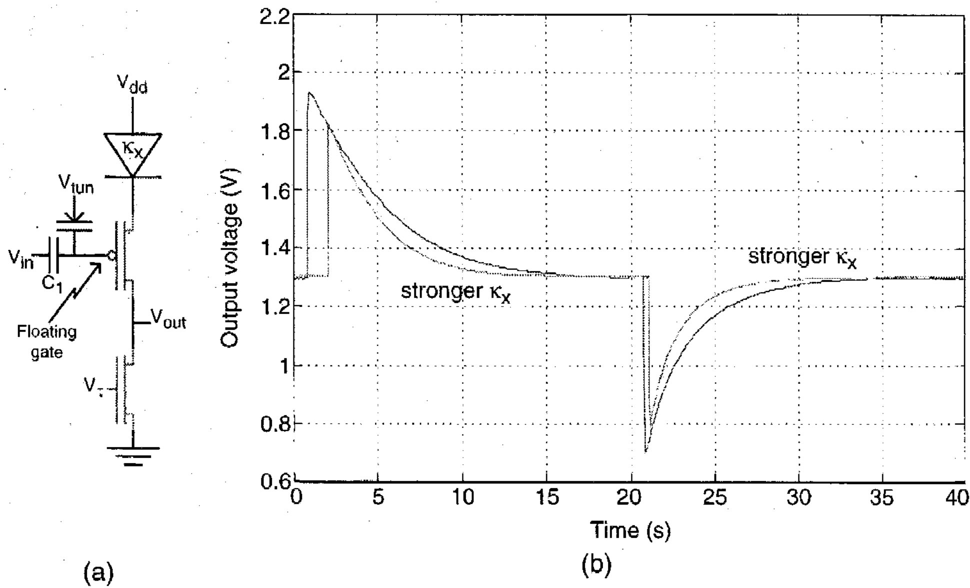


Figure 4.13: The behavior of the voltage autozeroing circuit using a source-degenerated pFET synapse. Unlike the nFET synapse, this circuit converges to its steady-state voltage. (a) Circuit diagram. (b) Response of this circuit to an upgoing and a downgoing step input for two different values of κ_x. The circuit behavior does not change for different channel currents if the bias drain–to–source voltage is fixed.

step changes in V_{in}, the output current converges toward the equilibrium current. This circuit is an autozeroing transconductance amplifier, because the output current always returns to the same equilibrium level.

4.2.3 Source-Degenerated pFET Floating-Gate Circuits

Figure 4.13 shows the source-degenerated pFET synapse with its drain connected to a current source. As in the AFGA case, the source and tunneling currents are nearly constant, and W is fixed at 1. Because the circuit returns to equilibrium by changing its drain voltage, the qualitative behavior is identical to that of the AFGA. The governing equation is

$$C_2 \frac{dV_{out}}{dt} = -C_1 \frac{dV_{in}}{dt} + I_{tun0} \left(\exp\left(\frac{\Delta V_{out}}{V_{inj}} \right) - 1 \right), \tag{4.13}$$

where the trajectories of this differential equation converge toward the steady state at

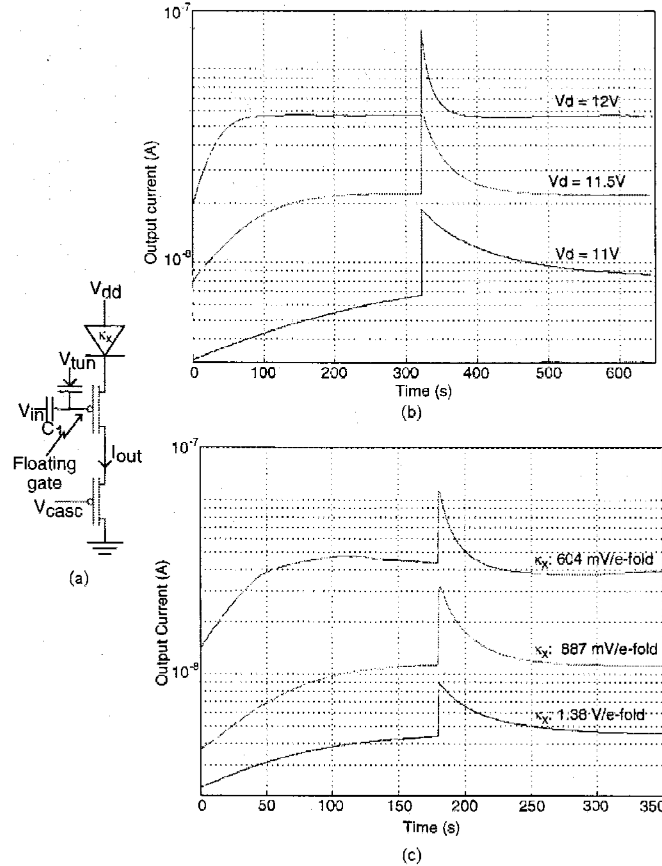


Figure 4.14: The behavior of the current autozeroing circuit using a source-degenerated pFET synapse. Unlike the pFET synapse, this circuit converges to its steady-state current. (a) Circuit diagram. (b) Response of this circuit to an upgoing and a downgoing step input for three values of drain voltage. (c) Response of this circuit to an upgoing and a downgoing step input for three different values of κ_x.

$\Delta V_{out} = 0$; this behavior is consistent with stabilizing negative feedback. The source voltage is almost constant for this low-frequency regime, because the source current is almost constant. Figures 4.13(b) shows the circuit response to input steps near the steady-state voltage; between the step changes in V_{in}, the output current converges toward its equilibrium voltage.

Figure 4.14 shows the source-degenerated pFET synapse with its drain connected to a cascode transistor; this configuration is also stable. For the current autozeroing configuration to be stable, the hot-electron injection current must decrease for an increasing change in the source-current; the stable behavior in both source-degenerated

circuits illustrates the changes in the source-current dependence on the hot-electron injection current from a pFET synapse. If the output current exceeds its equilibrium value ($W = 1$), then the injection current exceeds the tunneling current, causing both the floating-gate voltage and source voltage to decrease. For a sufficiently small κ_x, the change in the pFET's drain-to-source voltage decreases the injection current more than the amount by which the increasing source current increases the injection current. Decreasing the injection current increases the output current, and that, in turn, returns the output current back to equilibrium. If the output current is below its equilibrium value, then the tunneling current exceeds the injection current, increasing both the floating-gate voltage and the pFET's source voltage. For a sufficiently small κ_x, the change in drain-to-source voltage will increase the injection current more than the amount by which the decreasing source current decreases the injection current. Increasing the injection current decreases the output current, and that, in turn, returns the output current back to equilibrium. Because the drain voltage of the synapse is fixed as in Fig. 4.9, I obtain from (3.7) that

$$\frac{C_T U_T}{\kappa_n I_{tun0}} \frac{dW}{dt} = W \frac{C_1}{I_{tun0}} \frac{dV_{in}}{dt} + W^{1+\frac{U_T}{\kappa_x \kappa_p V_x}} - W^{1+\frac{\sigma}{\kappa_p \kappa_x}}; \tag{4.14}$$

the trajectories of this equation converge toward $W = 1$ and diverge away from $W = 0$, and are consistent with negative feedback. Figure 4.14 shows the circuit response to an input step near the steady-state voltage for various κ_x and drain voltages; between the step changes in V_{in}, the output current converges toward its equilibrium current. For the pFET and nFET synapses, only one of the two circuits is stable, but both source-degenerated pFET synapse configurations are stable. The positive r_{fg} stabilizes the cascoded-drain circuit, and the open-drain conductance, $r_{fg} + (g_m r_o) g_{fg}$, stabilizes the circuit when the drain is connected to a current source.

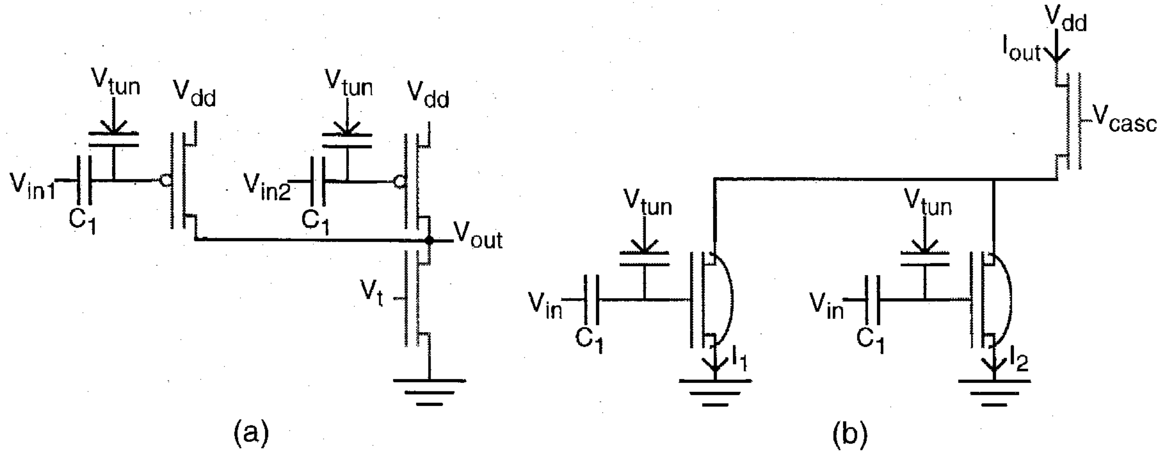


Figure 4.15: (a) Circuit with two pFET synapses coupled at the drain with a current source. (b) Circuit with two nFET synapses coupled at the drain with a cascode transistor.

Table 4.2: Values of A and B for the three synapses

Synapse	A	B
pFET	$W_1^{1+\alpha} + W_2^{1+\alpha}$	$W_1^{1-\frac{U_T}{\kappa_p V_x}} + W_2^{1+\frac{U_T}{\kappa_p V_x}}$
nFET	$W_1^{1+\alpha} + W_2^{1+\alpha}$	$W_1^{1-\frac{U_T}{\kappa_n V_x}} + W_2^{1-\frac{U_T}{\kappa_n V_x}}$
s-d pFET	$W_1^{1-\sigma} + W_2^{1-\sigma}$	$W_1^{1+\frac{U_T}{\kappa_x \kappa_p V_x}} + W_2^{1+\frac{U_T}{\kappa_x \kappa_p V_x}}$

4.3 Networks of Two Coupled Synapses

Where the previous section explored the possible behaviors of circuits comprising a single synapse transistor, this section considers the interaction between synapses coupled though their drain terminals. These networks show both competitive and cooperative behavior between synapses. I only use the stable circuit configurations from the previous section, since it is difficult to illustrate the behavior of circuits that have no stable operating point. I consider, in turn, the stable two-synapse circuits for the pFET, the nFET, and the source-degenerated pFET synapses. This section considers the circuit responses for a constant input voltage, because I want to analyze the relaxation of the synaptic currents.

Figures 4.15 and 4.16 show the multiple-synapse circuits that we are considering.

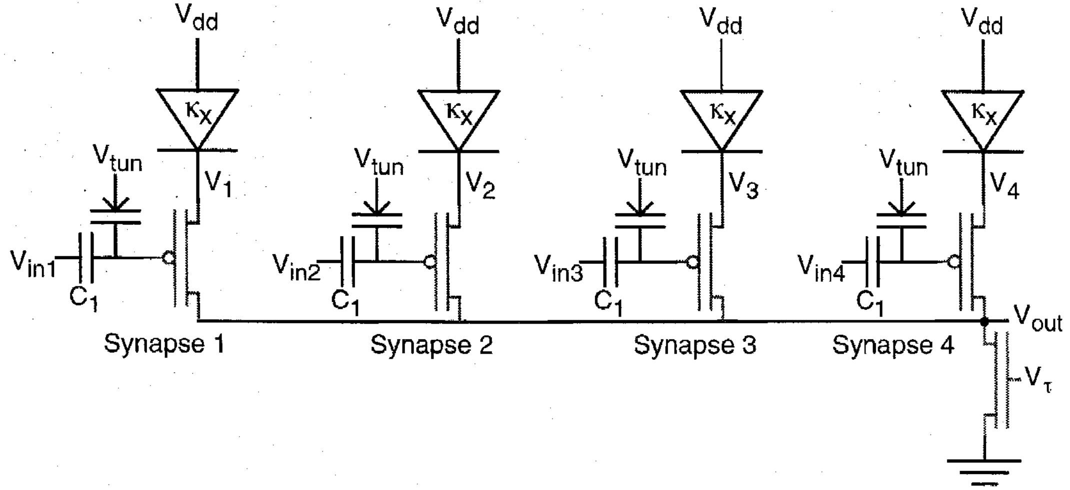


Figure 4.16: Circuit diagram of the four-input source-degenerated pFET synapse.

First, I want to show that the output voltages (V_{out}) and synapse currents ($I_{so}W$) return to their equilibrium values; the single- and multiple-synapse cases are similar in that V_{out} values return to their original steady states. The pFET synapse channel currents are constrained by a current source, which I define to be the sum of the bias currents in each synapse ($2I_{so}$); therefore, the sum of the two weights (W_1, W_2) is equal to 2. For a positive step into the gate of either synapse, V_{out} decreases, causing both transistors to inject more electrons onto their floating gates; consequently, the transistors source more current, and V_{out} increases. From the slow-variable definitions in (3.7), I model the output voltage for the traditional pFET synapse and the source-degenerated pFET synapse as

$$\frac{C_2}{I_{tun0}}\frac{dV_{out}}{dt} = Ae^{-\frac{\Delta V_{out}}{V_{inj}}} - B, \tag{4.15}$$

where A and B are functions of W_1, W_2, as defined in Table 4.3. A and B are always positive and vary slowly; as a result both equations are qualitatively identical to the output-voltage equation of the AFGA. The ΔV_{out} trajectories converge to the circuit's steady state. The nFET synapses are constrained by the cascode transistor; therefore, $\Delta V_{out} = -U_T \ln\left((W_1 + W_2)/2\right)$ for $C_2 = 0$. We differentiate this expression

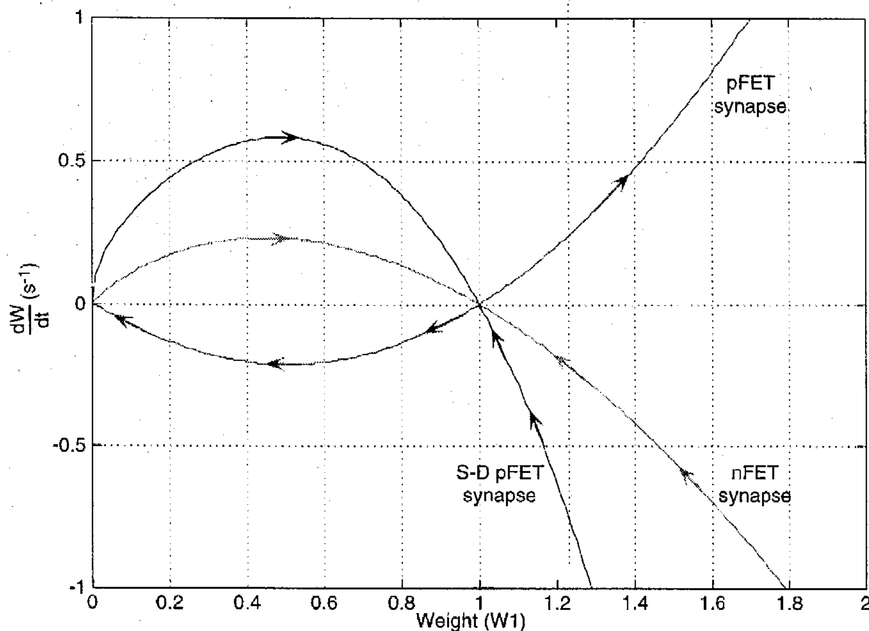


Figure 4.17: Plot of the time-derivative of W versus W for the nFET, pFET, and source-degenerated pFET synapses. The arrows show the directions that the differential equations will take. This figure shows that the nFET and source-degenerated synapses will stabilize to the $W = 1$ steady state, while the pFET synapse will diverge from the $W = 1$ steady state.

and substitute the slow-variable definitions in (3.7) to get the output voltage for the nFET synapse:

$$\frac{C_T + \kappa_n C_2}{\kappa_n I_{tun0}} \frac{dV_{out}}{dt} = B - Ae^{\frac{\Delta V_{out}}{V_{inj}}}. \tag{4.16}$$

Again A and B are positive, so the ΔV_{out} trajectories of (4.16) converge to the circuit's steady state, as in the pFET cases in (4.9) and (4.13).

Once the output voltage has reached equilibrium, the synapses act like coupled current autozeroing circuits; the behavior of the cascoded drain circuit for a particular synapse flavor determines the stability between synapses of that flavor. A constant output voltage requires that the sum of the synapses' weights be a constant. Because ΔV_{out} converges to 0 as $t \to \infty$, the stability of the weights is determined by the circuit behavior at $\Delta V_{out} = 0$. In practice, ΔV_{out} usually reaches its equilibrium before the weights reach their equilibria, and the weight adaptation does not affect

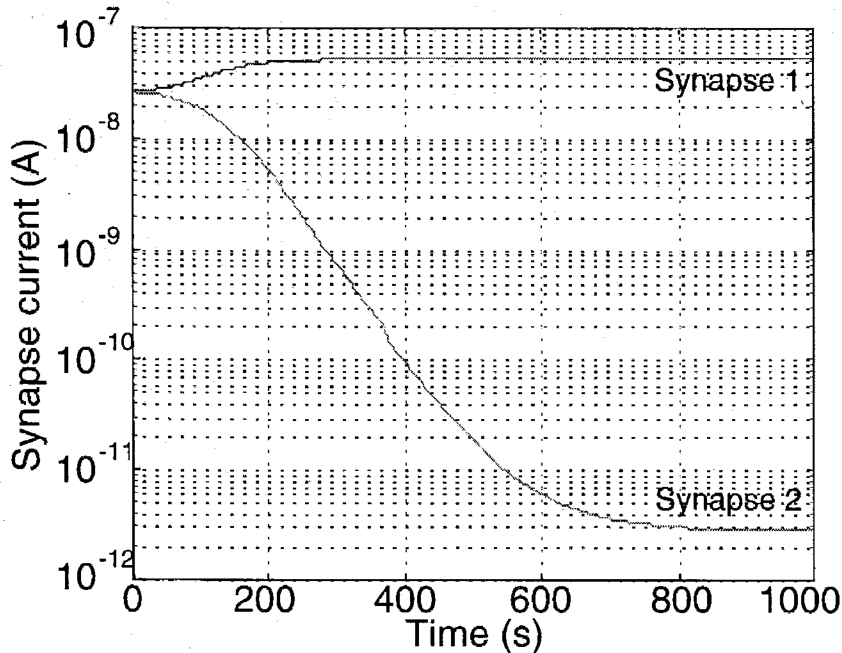


Figure 4.18: The behavior of coupled pFET synapses for fixed inputs. Even though the synapse currents initially start near each other, I_1 wins and I_2 loses. I_2 decreases as a linear exponential in time due to the constant tunneling current at the floating gate. The measured I_2 saturates due to the surrounding leakage currents; the floating gate continues to increase with time.

V_{out} once it has reached equilibrium. When ΔV_{out} has not converged to equilibrium, the behavior of the weights is similar qualitatively to the case when ΔV_{out} has reached equilibrium. In the following paragraphs, I discuss the stability of all three circuits; Fig. 4.17 illustrates qualitatively the stability of all three circuits by plotting $\frac{dW}{dt}$ versus W. I assume, without loss of generality, that $W_1 < W_2$, because the circuit behavior is symmetric in W_1 and W_2.

First, I consider the long-time behavior of the classic pFET synapse once $\Delta V_{out} = 0$. Figure 4.18 shows measured data of two synapse currents once $\Delta V_{out} = 0$. When W_2 is greater than W_1, the first synapse injects less than the second, further increasing the ratio of W_2 to W_1. Since the sum of the channel current is fixed, the first synapse takes more of the bias current and the current of the second synapse decreases; therefore, the second synapse's current steadily drops to zero. This behavior

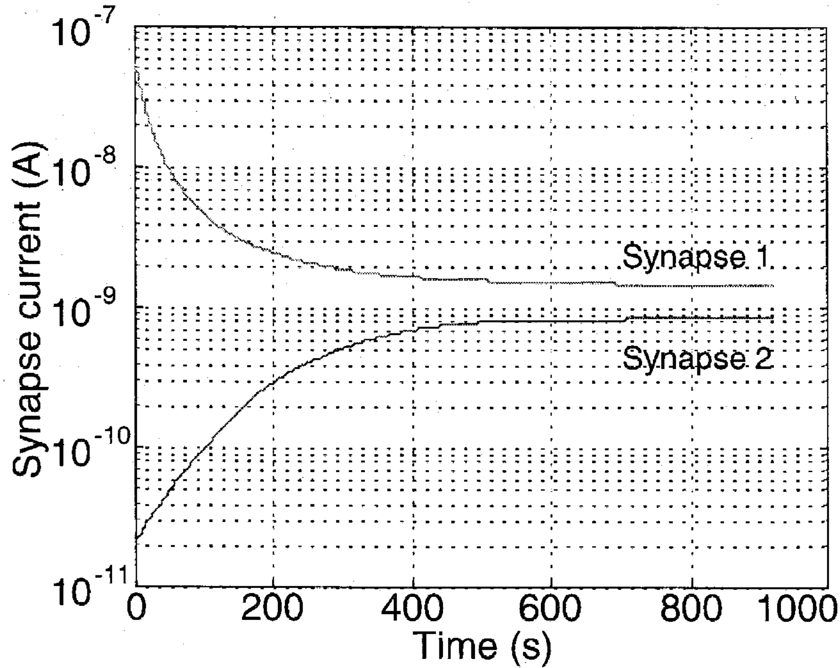


Figure 4.19: The behavior of coupled nFET synapses for fixed inputs. Even though the synapse currents initially start orders of magnitude apart from each other, both currents eventually converge to nearly the same steady-state level.

can be modeled for zero inputs as

$$\frac{C_T U_T}{\kappa_p I_{tun0}} \frac{dW_1}{dt} = W_1^{1+\frac{U_T}{\kappa_p V_x}} \left(W_1^{\alpha - \frac{U_T}{\kappa_p V_x}} - 1 \right). \tag{4.17}$$

Since I assumed that $W_2 > W_1$, the W_1 trajectories of this equation diverge from 1 and converge to 0. Figure 4.18 shows that, if the two starting weights are equal, over time one weight will decrease to 0 and the current in the other synapse will be equal to the bias current. If the losing weight is brought slightly above the winning weight, what was the losing synapse will now be the winner. The pFET synapses compete with each other for the bias current; in some sense, this circuit displays winner-take-all behavior in the weight space. This behavior is typical of a continuous-time normalizing Hebbian network; after a period of time, we cannot reuse these synapses without significantly altering this circuit.

Next, I consider the long-time behavior of the nFET synapses once $\Delta V_{out} = 0$.

The two nFET synapses are coupled with a cascode connection to the common drain terminal, which pins the drain voltage. A nearly fixed drain voltage means that the synapses are only weakly coupled through the drain; I could achieve stronger coupling with a negative-resistance circuit. If W_1 is smaller than W_2, the first synapse injects less than the second, differentially increasing the floating-gate voltages, and resulting in a larger W_1 and smaller W_2. I model the nFET synapse behavior as

$$\frac{C_T U_T}{\kappa_n I_{tun0}} \frac{dW_1}{dt} = W_1^{1-\frac{U_T}{\kappa_n V_x}} \left(1 - W_1^{1+\frac{U_T}{\kappa_n (V_x || - V_{inj})}}\right). \tag{4.18}$$

The W_1 trajectories of this equation diverge from 0 and converge to 1. Because the sum of the channel current is fixed, the first synapse takes a larger fraction of the total current. The weights of the two synapses converge to the steady state $W_1 = W_2$. Figure 4.19 shows that, if the two starting weights are different, then, over time, the two weights will converge to the same value.

Third, I consider the long-time behavior of the source-degenerated pFET synapse once $\Delta V_{out} = 0$. I model the source-degenerated pFET synapse for constant inputs as

$$\frac{C_T U_T}{\kappa_p \kappa_x I_{tun0}} \frac{dW_1}{dt} = W_1^{1-\frac{\sigma}{\kappa_p \kappa_x}} \left(1 - W_1^{\frac{U_T}{\kappa_x \kappa_p V_x} + \frac{\sigma}{\kappa_p \kappa_x}}\right). \tag{4.19}$$

Since I assumed that $W_2 > W_1$, the W_1 trajectories of this equation diverge from 0 and converge to 1. This circuit shows stablizing feedback between the two synapses, because the current autozeroing configuration for a single synapse was stable. Also, the threshold for stability is the point at which the exponent of the second term in W_1 is equal to 0; this condition can be expressed as

$$\kappa_x = \frac{U_T}{\kappa_p V_x || V_{inj}}, \tag{4.20}$$

for small κ_x. Figure 4.20 shows the stable response of a four-input source-degenerated pFET synapse network. Since all synapses of the four-synapse circuit converge to equilibrium, we can see one example that these synapse properties extend to multiple synapses.

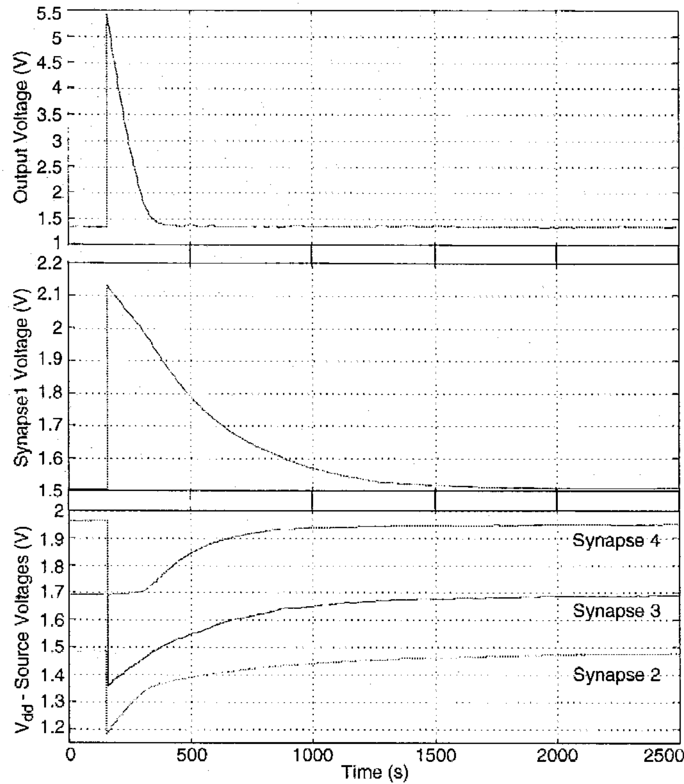


Figure 4.20: Output voltage and synapse voltage (V_1, V_2, V_3, V_4) responses due to an input step applied to the first synapse. This figure shows that the stabilizing behavior for two synapses is extendible to multiple synapses.

4.4 Conclusions

I have characterized and modeled the dynamics of nFET and pFET single-transistor synapses operating in continuous-time circuits. The dynamic behavior of a single synapse can be characterized from that synapse's response in both a constant-current configuration and a constant-voltage configuration. For a pFET synapse, the constant-current configuration is stable, because the floating-gate feedback from its drain is stable. For an nFET synapse, the constant-voltage configuration is stable, because the floating-gate feedback from its floating gate is stable. The pFET synapse's constant-voltage configuration and the nFET-synapse's constant-current configuration are unstable circuits. I presented a new type of pFET synapse where degenerating the source results in gate currents providing stabilizing feedback in both

configurations.

I characterized and modeled the dynamics of two coupled single-transistor synapses, where the sum of the weights converges to a steady state. The two pFET synapses compete for the available bias current; the synapse starting with the larger channel current will eventually supply all the bias current. The two nFET synapses cooperate for the entire available channel current; regardless of the starting position, the two synapses converge to nearly equal channel currents. Since the gate currents of the source-degenerated pFET synapse provide stabilizing feedback to the floating-gate and drain, the weights of coupled source-degenerated pFET synapses converge to nearly equal channel currents, as they do in the nFET synapses. These properties extend directly to multiple synapses.

Chapter 5 An Autozeroing Floating-Gate Amplifier

This chapter presents a bandpass floating-gate amplifier that uses tunneling and pFET hot-electron injection so that it can return to its sensitive region despite large changes in the DC input voltage. Offsets often present a difficult problem for designers of MOS analog circuits. A time-honored tradition for addressing this problem is to use a blocking capacitor to eliminate the input DC component; however, for integrated filters, this approach requires enormous input capacitors and resistors to get time constants of less than 1Hz. Existing on-chip autozeroing techniques rely on clocking schemes that compute the input offset periodically, then subtract the correction from the input [85]. These autozeroing techniques add significant complexity to the circuit, as well as clock noise, aliasing, etc.

This chapter presents the *autozeroing floating-gate amplifier* (AFGA). The AFGA is an integrated continuous-time filter that is intrinsically autozeroing. It can achieve a high-pass characteristic at frequencies well below 1Hz. In contrast with conventional autozeroing amplifiers that eliminate their input offset, the AFGA nulls its output offset. The AFGA is a continuous-time filter; it does not require any clocking. The AFGA is the first known application of pFET hot-electron injection. Until now, pFET hot-electron injection has attracted attention only as a source of MOSFET oxide degradation [86]. The autozeroing technique used in the AFGA can be applied to a wide variety of floating-gate MOS circuits (FGMOS) to continuously restore a desired baseline operation on a slow timescale.

Section 5.1 gives a qualitative overview of AFGA operation. Section 5.2 considers how the steady-state voltage varies with the circuit parameters—most notably with the tunneling voltage and the nFET bias current. Section 5.3 considers the AFGA's high-pass filter behavior, and addresses long-term parameter drift. Section

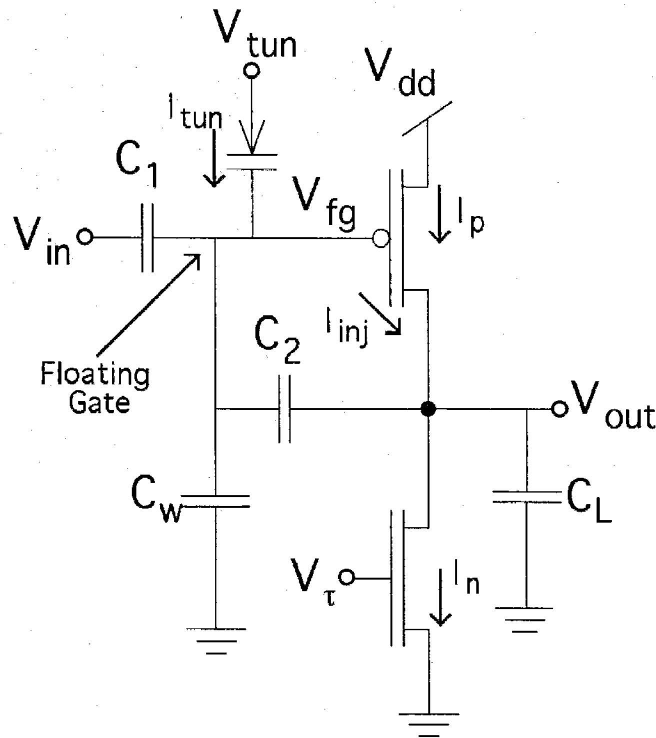


Figure 5.1: An autozeroing floating-gate amplifier (AFGA) that uses *p*FET hot-electron injection. The ratio of C_2 to C_1 sets the gain of this inverting amplifier. The *n*FET is a current source, and it sets the current through the *p*FET. Steady state occurs when the injection current is equal to the tunneling current. The capacitance from the floating gate to ground, C_w, represents both the parasitic and the explicitly drawn capacitances. Increasing C_w will increase the linear input range of the circuit. The capacitance connected to the output terminal, C_L, is the load capacitance. Between V_{tun} and V_{fg} is our symbol for a tunneling junction, which is a capacitor between the floating-gate and an *n*well.

5.4 considers the AFGA's low-pass filter behavior. Section 5.5 describes the AFGA's frequency response and dynamic range. Section 5.6 investigates how the steady-state voltage changes with input signal amplitude. Up to that point, this chapter will have primarily considered the AFGA biased with subthreshold currents; the discussion in Section 5.7 considers the differences in circuit operation when the AFGA is biased with above-threshold currents. Section 5.7 also discusses other AFGA effects. I will present some conclusions Section 5.8.

5.1 Qualitative Presentation of AFGA Operation

Figure 5.1 shows the autozeroing floating-gate amplifier. The open-loop amplifier consists of a pFET input transistor and an nFET current source. With capacitive feedback, the input signal is amplified by a closed-loop gain approximately equal to $-\frac{C_1}{C_2}$. The maximum gain is limited both by the open-loop gain, and by the parasitic floating-gate–to–drain overlap capacitance.

The complementary tunneling and hot-electron injection processes adjust the floating-gate charge such that the amplifier's output voltage returns to a steady-state value on a slow time scale. If the output voltage is below its equilibrium value, then the injection current exceeds the tunneling current, decreasing the charge on the floating gate; that, in turn, increases the output voltage back toward its equilibrium value. If the output voltage is above its equilibrium value, then the tunneling current exceeds the injection current, increasing the charge on the floating gate; that, in turn, decreases the output voltage back toward its equilibrium value. The circuit behaves like a high-pass filter with a long time constant.

Two conditions must be satisfied for the circuit to be in equilibrium. First, the pFET channel current, I_p, must be equal to the nFET channel current, I_n. I shall define this quiescent channel current as I_{s0}. Second, the injection gate current must be equal to the tunneling gate current. I shall define I_{inj0} as the quiescent injection current that must equal I_{tun0}, the quiescent tunneling current, at equilibrium. Since the tunneling and injection currents are many orders of magnitude smaller than I_{so} and are charging similar-sized capacitances, the first condition is satisfied much faster than is the second condition. The frequency range over which the first condition is satisfied, but the second condition is not satisfied, is where the AFGA behaves as an amplifier. The combination of electron tunneling and pFET hot-electron injection applies the appropriate negative feedback to stabilize the output voltage such that the second condition also is satisfied.

In the frequency range where the first condition does not hold, the output voltage is attenuated. In this regime, the circuit behaves as a low-pass filter. Since the

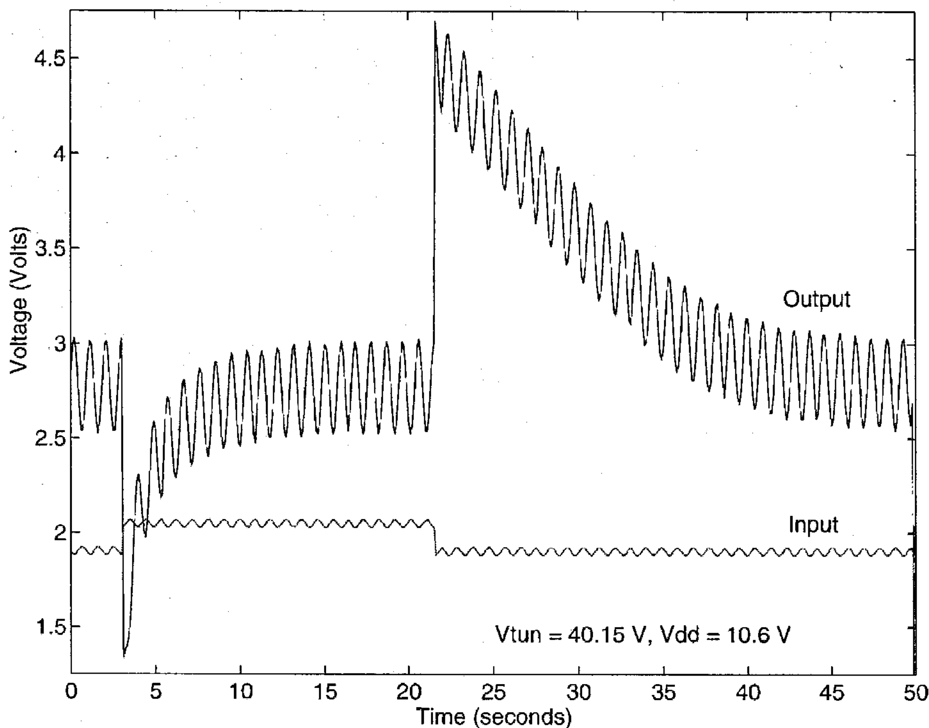


Figure 5.2: Response of the AFGA to a 1Hz sinewave superimposed on a 19s voltage pulse. The AFGA has a closed-loop gain of 11.2, and a low-frequency cutoff at 100mHz. The 1Hz signal is amplified, but the much slower step is adapted away.

output capacitances are charged or discharged by currents on the scale of I_{s0}, the cutoff frequency will be directly dependent on the bias current. Continuous-time integrators operate on a similar principle [87, 82]. The AFGA transfer function is a bandpass, with the low-frequency cutoff set by the equilibrium tunneling and injection currents, and the high-pass cutoff independently set by the equilibrium pFET and nFET channel currents.

Figure 5.2 shows the response of the autozeroing floating-gate amplifier to a 1Hz sine wave superimposed on an input pulse. If the input changes on a timescale that is much shorter than the adaptation, then the output is an amplified version of the input signal. The amplifier adapts to the pulse input after an initial transient, while preserving the amplified 1Hz sine wave.

Before proceeding, let us consider the behavior of the Early voltage, because

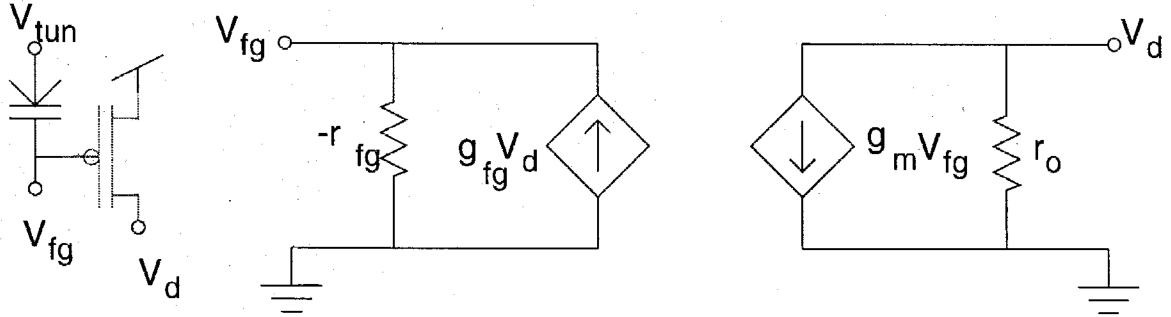


Figure 5.3: The small-signal model of a pFET with the effects of hot-electron injection. I assume a constant tunneling current at the floating gate (V_{fg}); this tunneling current sets the bias point for the hot-electron injection parameters.

the Early voltage is directly related to the amplifier's open-loop gain. Figure 5.3 reviews the small-signal model of a pFET that includes the effects of tunneling and hot-electron injection. Following the conventional definitions of small-signal transconductance, g_m, and output resistance, r_o [81], we obtain

$$g_m = \frac{\kappa I_{so}}{U_T}, \quad \text{and} \quad r_o = \frac{V_o}{I_{so}}, \tag{5.1}$$

for a subthreshold pFET. Figure 5.4 shows how the Early voltage, V_o, changes when the FETs operate with large drain-to-source voltages. The Early voltage decreases at large drain-to-source voltages due to impact ionization in the drain-to-channel depletion region. In the drain-to-channel depletion region, holes are accelerated to large energies; if a hole has an energy larger than the bandgap, then it may undergo impact ionization. The result of an impact ionization is two holes and one electron. For the nFET biased with a drain-to-source voltage of 3.0V and the pFET biased with a drain-to-source voltage of 8.5V, V_o is nearly constant for both transistors; therefore the AFGA's open-loop gain is also nearly constant. For the amplifiers presented in this chapter, the maximum open-loop gain is roughly 1400. The definitions of the small-signal quantities, g_{fg} and r_{fg}, were presented in Chapter 4.

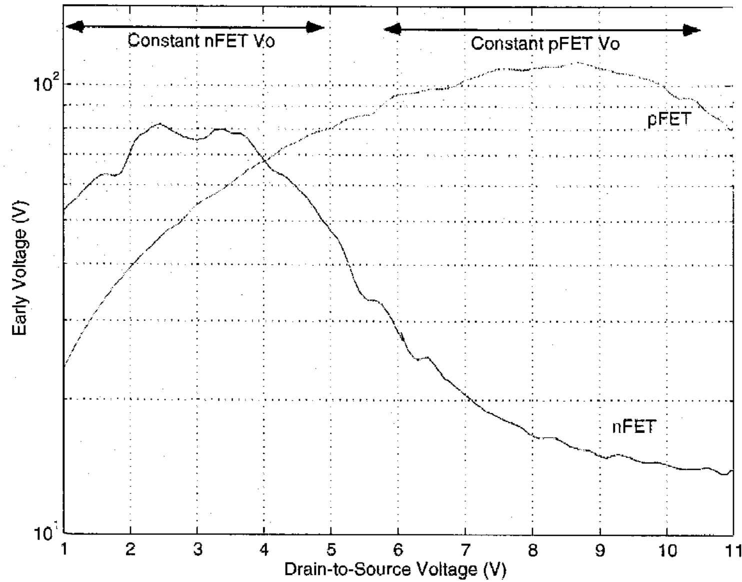


Figure 5.4: The effect of drain-to-source voltage on the Early voltage of an nFET and pFET. The AFGA's open loop-gain as a function of the output voltage is directly related to the change in the Early voltage. The decrease in the nFET's Early voltage at high drain-to-source voltages is due to the impact ionization in its drain-to-channel depletion region. For typical steady-state output voltages around 1V to 5V, the Early voltage of both the pFET and nFET are nearly constant; therefore, the open-loop gain is constant.

5.2 Equilibrium Voltages of the AFGA

Qualitatively, two factors change the steady-state output voltage. For the injection current to match the tunneling current after a change in V_T or V_{tun}, the output voltage must reach a new equilibrium. Increasing the bias voltage or channel current requires increasing in the output voltage, because the pFET must reduce its injection efficiency so that the injection current matches the original tunneling current. Increasing the tunneling voltage, which increases the steady-state tunneling current, requires decreasing the output voltage, because the pFET must increase its injection efficiency so that the injection current matches the new tunneling current.

In Section 5.1, I postulated two conditions for equilibrium; now I describe them quantitatively. I assume an initial operating point, and consider changes in the steady-state output voltage in response to a change in V_T or in V_{tun}, as in Chapter 3. First,

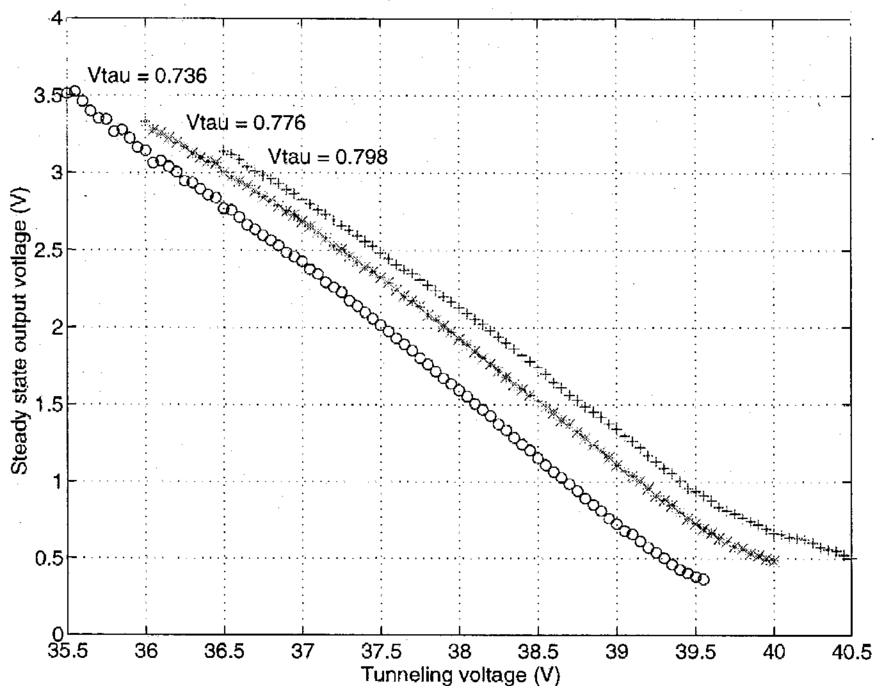


Figure 5.5: Steady-state output voltage versus the tunneling voltage for three values of V_τ. This data set agrees with the model described in (5.5). The AC gain of this amplifier was 146.

the current in the nFET must be equal to the current in the pFET:

$$\begin{aligned} I_n &= I_p, \\ I_{so} \exp\left(\frac{\kappa_n \Delta V_\tau}{U_T}\right) &= I_{so} \exp\left(\frac{-\kappa \Delta V_{fg}}{U_T}\right), \end{aligned} \tag{5.2}$$

where ΔV_τ is the change in the bias voltage, and ΔV_{fg} is the change in the pFET's floating-gate voltage. Therefore, we can solve for the relation

$$\Delta V_{fg} = -\frac{\kappa_n}{\kappa} \Delta V_\tau. \tag{5.3}$$

Qualitatively, the bias current in the nFET sets the current in the pFET, and therefore sets the floating-gate voltage.

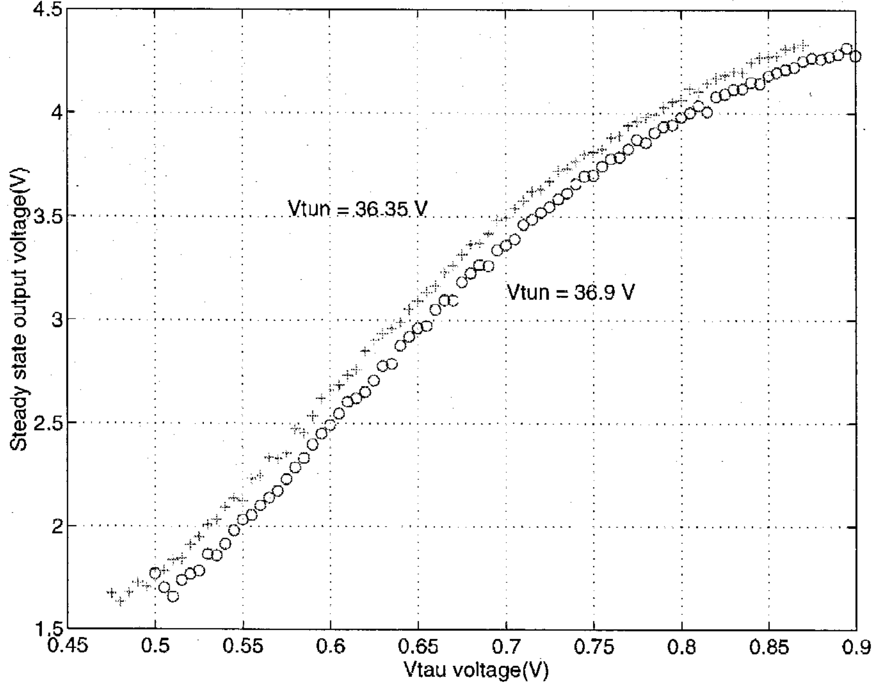


Figure 5.6: Steady-state output voltage versus V_τ for two tunneling voltages. This data set agrees with the model described in (5.5). The AC gain of this amplifier was 146.

Second, the tunneling current must be equal to the injection current:

$$I_{inj0} e^{-\frac{\alpha\kappa\Delta V_{fg}}{U_T}} e^{\frac{-\Delta V_{out}}{V_{inj}}} = I_{tun0} e^{\frac{\Delta V_{tun} - \Delta V_{fg}}{V_x}}. \tag{5.4}$$

I assume that I_{inj0} is equal to I_{tun0} at the initial operating point. Then, we obtain the second equilibrium relationship:

$$\Delta V_{out} = -\frac{V_{inj}}{V_x}\Delta V_{tun} + V_{inj}\frac{\kappa_n}{\kappa}\left(\frac{\alpha\kappa}{U_T} - \frac{1}{V_x}\right)\Delta V_\tau. \tag{5.5}$$

For above-threshold operation, the above equation becomes

$$\Delta V_{out} = -\frac{V_{inj}}{V_x}\Delta V_{tun} + V_{inj}\frac{\kappa_n}{\kappa}\left(\frac{2}{V_{dd} - V_{fg} + V_T} - \frac{1}{V_x}\right)\Delta V_\tau. \tag{5.6}$$

Figures 5.5 and 5.6 show the measured equilibrium output voltage as a function of the bias parameters for an AFGA with a gain of 146. Figure 5.5 shows the equilibrium

output voltage versus the tunneling voltage. Figure 5.6 shows the equilibrium output voltage versus the bias current; the curves begin to saturate for above-threshold bias currents, as predicted by (5.6).

An amplifier should be insensitive to variations in the bias voltages; the data in Figs. 5.5 and 5.6 show the circuit's DC sensitivity to V_τ and V_{tun}. First, the DC gain from the tunneling node to the output is given by $\frac{V_{inj}}{V_x}$, which, for this measured data, is 0.64. The DC gain from the nFET gate to the output is given by $V_{inj}\frac{\kappa_n}{\kappa}\left(\frac{\alpha\kappa}{U_T} - \frac{1}{V_x}\right)$, which, for this measured data, is 10. The DC gain from the input to the output is zero. All three DC gains are smaller than the AFGA AC gain of 146.

The power-supply rejection depends on the choice of reference. With the input and output referenced to V_{dd}, and with V_τ referred to GND, the AFGA gain from the power-supply to the output is 0.1, which results in a power-supply rejection ratio in the passband of 64dB. The power-supply rejection ratio is limited by the open-loop gain of the amplifier. If we instead refer all nodes to GND, the AC power supply gain is 146, which decreases the power-supply rejection ratio to 0dB. The circuit designer must also be careful of where the floating-gate capacitances are connected; for good power-supply rejection, all these capacitors must be referenced to V_{dd}. Otherwise, the power supply becomes another input to the AFGA, and any AC power supply noise will appear at the output, amplified by the AFGA's AC gain.

The steady-state output voltage and the high-pass cutoff frequency are set explicitly by V_{tun} and by the power-supply voltage. Increasing V_{tun} increases the tunneling current, which in turn decreases the settling time, but also decreases the steady-state output voltage, since the pFET must increase its drain-to-source voltage. Increasing the power-supply voltage decreases the tunneling current by decreasing the voltage across the oxide, increasing the settling time and increasing the steady-state output voltage.

5.3 Low-Frequency AFGA Behavior

The quantitative AFGA dynamics are described by two general equations governing the autozeroing floating-gate amplifier around an equilibrium output voltage. We can obtain the first equation by applying Kirchoff's current law (KCL) at the floating gate:

$$(C_1 + C_2 + C_w)\frac{dV_{fg}}{dt} = C_1\frac{dV_{in}}{dt} + C_2\frac{dV_{out}}{dt} + I_{tun0}\left(1 - \exp\left(-\alpha\frac{\kappa\Delta V_{fg}}{U_T} - \frac{\Delta V_{out}}{V_{inj}}\right)\right). \tag{5.7}$$

We can obtain the second equation by applying KCL at the output node:

$$(C_2 + C_L)\frac{dV_{out}}{dt} = C_2\frac{dV_{fg}}{dt} + I_\tau\left(\exp\left(-\frac{\kappa\Delta V_{fg}}{U_T}\right) - 1\right). \tag{5.8}$$

I have neglected the Early effect, which adds a correction term to (5.8). As long as the closed-loop gain is much lower than the amplifier gain, ignoring the Early effect is a good approximation.

In the passband, where the AFGA is an amplifier, the floating gate is held nearly constant by the amplifier feedback, and the tunneling and injection currents are negligible. This approximation simplifies (5.7) to

$$C_2\frac{dV_{out}}{dt} = -C_1\frac{dV_{in}}{dt}; \tag{5.9}$$

thus, the change in the output voltage (ΔV_{out}) is equal to the input voltage (ΔV_{in}) amplified by $-\frac{C_1}{C_2}$.

5.3.1 Low-Frequency Model

I make an approximation to model the low-frequency response of the AFGA. The open-loop gain from the floating gate to the output can be large. If the output voltage is to be kept between the supply rails, the floating-gate voltage must be confined to a 10mV swing. Thus, I approximate the floating-gate voltage to be constant. Therefore, because the floating-gate voltage is nearly constant, the source

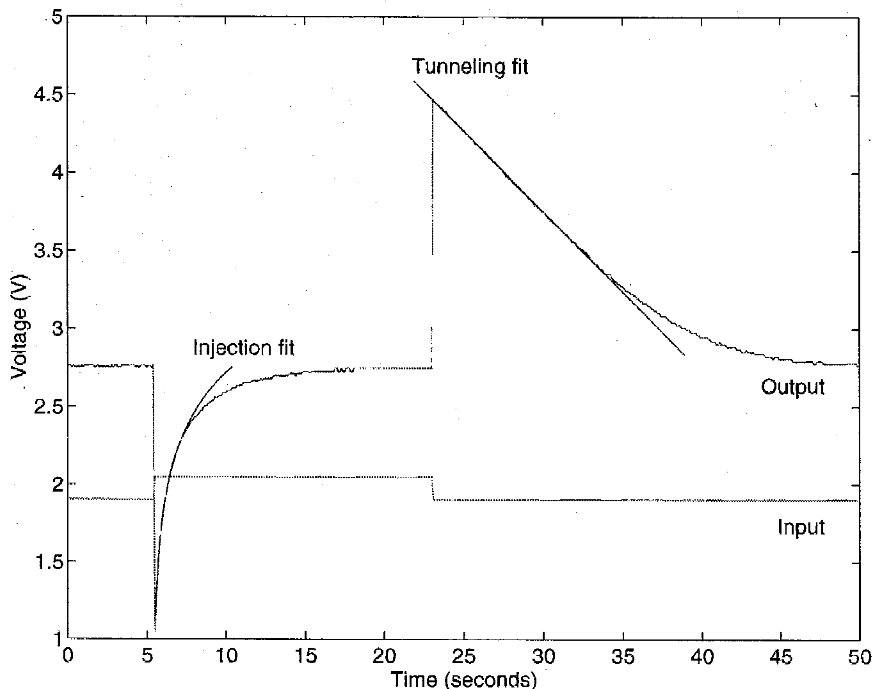


Figure 5.7: Response of the AFGA to an upgoing and a downgoing step input. The adaptation in response to an upward step results from electron tunneling; the adaptation in response to a downward step results from pFET hot-electron injection. This amplifier had a gain of 11.2. I plot the curve fits of the simplified expressions of (5.17), where either tunneling or injection dominates the restoration process. Using the curve fits, τ is 4.3s and I_{tun0} is 50fA. The value of τ can be set reliably to more than 10^5 seconds.

current varies only slightly. The quiescent source current (I_{s0}) is set by the nFET current source. From (3.17), the model of injection current for a fixed a source current I_{s0} is

$$I_{tun} - I_{inj} = I_{tun0}\left(1 - \exp\left(-\frac{\Delta V_{out}}{V_{inj}}\right)\right), \tag{5.10}$$

where $I_{tun0} = I_{inj0}$ for the circuit in equilibrium. Since the floating gate is held nearly constant by feedback, the floating-gate voltage dependence in (3.17) is negligible. Even when the circuit is biased with above-threshold currents, the tunneling current still remains nearly constant. Since the injection efficiency is still an exponential function of the drain voltage for above-threshold currents, the low-frequency dynamics are similar in below- and above-threshold operation.

With the preceding approximations, we can model the amplifier's output voltage,

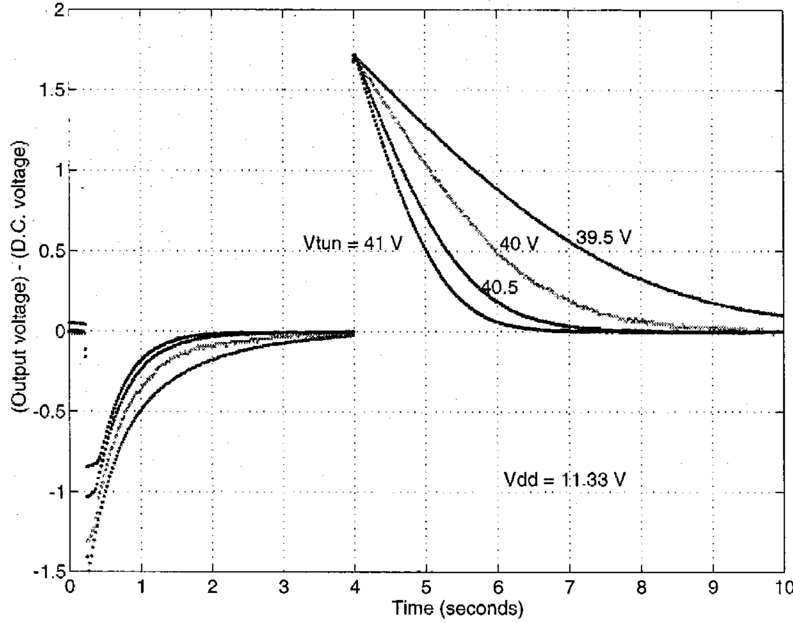


Figure 5.8: The response to a square wave for four different values of the tunneling voltage. This amplifier had a gain of 147; the input square wave is not shown. The steady-state output voltage decreased in the same manner as seen in Fig. 5.5 for increasing tunneling voltages. The initial tail in the upgoing response is due to the output voltage going to ground.

V_{out}, in terms of V_{in}, with a single equation. The total floating-gate current is the sum of the capacitive currents of the input and output terminals, plus the tunneling and injection currents. From (5.7) I write

$$C_2 \frac{dV_{out}}{dt} = -C_1 \frac{dV_{in}}{dt} + I_{tun0} \left(\exp\left(-\frac{\Delta V_{out}}{V_{inj}} \right) - 1 \right). \tag{5.11}$$

To solve (5.11), I make the following change of variables:

$$X = e^{\frac{\Delta V_{out}}{V_{inj}}}. \tag{5.12}$$

The resulting equation for X is a linear, first-order differential equation with variable coefficients

$$\tau_l \frac{dX}{dt} = -\frac{\tau_l A_v X}{V_{inj}} \frac{dV_{in}}{dt} + 1 - X, \tag{5.13}$$

where τ_l, the low-frequency cutoff, is equal to $\frac{C_2 V_{inj}}{I_{tun0}}$, and A_v is the closed-loop AC gain of the amplifier, $-\frac{C_1}{C_2}$.

5.3.2 Response to a Voltage Step

Consider the AFGA's response to an input voltage step. Assume that the output voltage initially has adapted to its steady-state value. To solve (5.13), I first assume that the output voltage immediately after applying the step, $\Delta V_{out}(0^+)$, is given by the magnitude of the input step times the AFGA AC gain. I employ $\Delta V_{out}(0^+)$ as a new effective initial condition, and denote the effective initial condition in X by

$$X(0^+) = \exp\left(\frac{V_{out}(0^+)}{V_{inj}}\right). \tag{5.14}$$

For a downward step, $X(0^+)$ is greater than 1; for an upward step, $X(0^+)$ is less than 1. After the input step, $\frac{dV_{in}}{dt} = 0$; therefore (5.13) becomes

$$\begin{cases} \tau_l \frac{dX}{dt} = 1 - X \\ X(0) = X(0^+). \end{cases} \tag{5.15}$$

The solution to (5.15) in terms of ΔV_{out} is

$$\Delta V_{out}(t) = V_{inj} \ln\left(1 + \left(X(0^+) - 1\right) e^{-\frac{t}{\tau_l}}\right), \tag{5.16}$$

where $\Delta V_{out} \to 0$ as $t \to \infty$.

The step response has three interesting regimes, which are approximated by

$$\Delta V_{out} \approx \begin{cases} \Delta V_{out}(0^+) e^{-\frac{t}{\tau_l}}, & X(0^+) \approx 1, \\ \Delta V_{out}(0^+) - \frac{I_{tun0}}{C_2} t, & X(0^+) \gg 1, \\ V_{inj} \ln\left(X(0^+) + \frac{t}{\tau_l}\right), & X(0^+) \ll 1. \end{cases} \tag{5.17}$$

The first case occurs when the tunneling current is nearly equal to the injection current just after the voltage step. The solution in this region is the familiar exponential decay

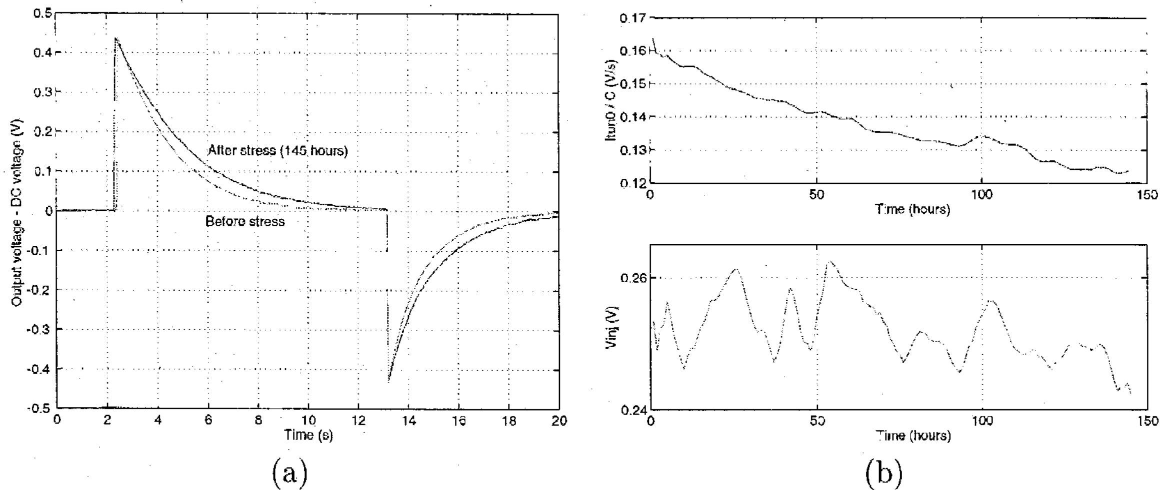
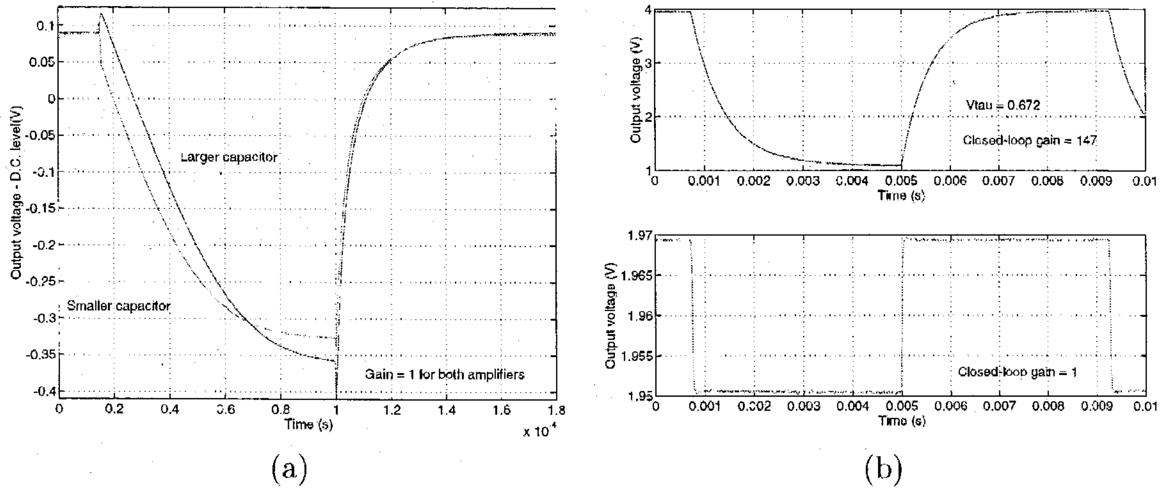


Figure 5.9: The effect of long-duration AFGA operation. (a) The responses to an upgoing and downgoing voltage step before and after 145 hours of operation. I plot the difference in the output voltage from the equilibrium DC level as a function of time; the equilibrium output voltage increased slightly over the 145 hours of operation. (b) The extracted device parameters as a function of time. Since I_{tun0}/C_2 changes more than does V_{inj}, most of the long-term change is caused from the tunneling junction, which is probably caused by oxide trapping.

of a linear system. The second case occurs when the tunneling current dominates the injection current. The behavior of the output voltage in this regime results from the constant tunneling current removing electrons from the floating gate. The third case occurs when the injection current dominates the tunneling current. Figure 5.7 shows a measured response to an input pulse, with curve fits to the regions where either the tunneling or injection current dominates.

5.3.3 Long-Term Parameter Drift

The physical properties of the tunneling and hot-electron injection mechanisms change with time. These processes are permanently modified as electrons pass through the oxide, creating electron traps. I investigated the long-term changes by performing an accelerated stress experiment, where I operated an AFGA continuously for 145 hours with an average τ_l of 1.7s. When an AFGA is used as an amplifier or as a



(a) (b)

Figure 5.10: High-frequency AFGA behavior. (a) Two AFGAs with unity gain, but with different values for C_1. The larger-capacitor circuit had $C_1 = C_2 = 300$ fF; the smaller-capacitor circuit had $C_1 = C_2 = 50$ fF. For both AFGAs, C_L was the same. I operated the two AFGAs with different subthreshold bias currents to achieve comparable settling times. (b) Two AFGAs with different gains.

low-pass filter, a more reasonable τ_l would be at least several minutes; therefore, this experiment represents the stress of operating the AFGA continuously for a few years. The effect of an input signal only slightly modifies the results of this experiment. To characterize the behavior of the AFGA over time, I performed a square wave experiment similar to the one shown in Fig. 5.7, once per hour for 145 hours. To each of the resulting output waveforms, I fit the expressions of (5.17) and extracted the relevant device parameters. Figure 5.9a shows the square-wave response of the AFGA before and after this lifetime test. The adaptation time constant has increased noticeably, but the general behavior is unaffected. Figure 5.9b shows the change in the extracted parameters during the experiment. Since I_{tun0}/C_2 changes more than V_{inj}, we see that most of the long-term change is due to changes in the tunneling junction, which is probably a consequence of electron trapping.

5.4 High-Frequency AFGA Behavior

For sufficiently high frequencies, the autozeroing floating-gate amplifier is a low-pass filter. In this regime, the tunneling and injection currents are negligible; therefore we approximate (5.7) as

$$(C_1 + C_2 + C_w)\frac{dV_{fg}}{dt} = C_1\frac{dV_{in}}{dt} + C_2\frac{dV_{out}}{dt}. \tag{5.18}$$

From (5.18), we see that changes in V_{out} are proportional to changes in V_{fg} and V_{in}. At extremely high frequencies, the transistor channel currents are negligible compared to the capacitive currents. In this capacitive-feedthrough regime, the solutions to (5.8) and (5.18) are

$$\begin{aligned}\frac{\Delta V_{fg}}{\Delta V_{in}} &= \frac{C_1(C_2+C_L)}{(C_1+C_2+C_w)(C_2+C_L)-C_2^2}, \\ \frac{\Delta V_{out}}{\Delta V_{in}} &= \frac{C_1 C_2}{(C_1+C_2+C_w)(C_2+C_L)-C_2^2}.\end{aligned} \tag{5.19}$$

We can reduce the effects of the capacitive feedthrough by increasing either C_L or C_W.

At frequencies between the low-frequency cutoff and the capacitive-feedthrough regime, the behavior of the AFGA results from the floating-gate voltage settling back to its equilibrium value. Therefore, we can combine (5.8) and (5.18) to form a single equation for the floating-gate voltage, which we write as

$$\begin{aligned}&\left((C_1 + C_2 + C_w)(C_2 + C_L) - C_2^2\right)\tfrac{dV_{fg}}{dt} = \\ &C_1(C_2 + C_L)\tfrac{dV_{in}}{dt} + C_2 I_\tau\left(e^{-\frac{\kappa\Delta V_{fg}}{U_T}} - 1\right).\end{aligned} \tag{5.20}$$

This equation is similar to (5.13), which describes the output-voltage response in the low-frequency case. Correspondingly, substituting

$$Y = e^{\frac{\kappa\Delta V_{fg}}{U_T}} \tag{5.21}$$

into (5.20) results in the linear differential equation

$$\tau_h \frac{dY}{dt} = \frac{\tau_{h2}\kappa}{U_T}\frac{dV_{in}}{dt} Y + 1 - Y, \tag{5.22}$$

where I shall define τ_{h2} to be

$$\tau_{h2} = \frac{C_1(C_2 + C_L)U_T}{\kappa C_2 I_\tau}, \tag{5.23}$$

which is the time constant that marks the onset of capacitive feedthrough. I shall define τ_h to be

$$\tau_h = \frac{\left((C_1 + C_2 + C_w)(C_2 + C_L) - C_2^2\right) U_T}{\kappa C_2 I_\tau}, \tag{5.24}$$

which represents the time constant for the high-frequency cutoff.

As we did in the low-frequency case, we shall consider the response to an input voltage step. To solve (5.22), I first assume that the floating-gate voltage immediately after applying the step, $\Delta V_{fg}(0^+)$, is given by the magnitude of the input step attenuated by the capacitive divider ratio, (5.19). With this initial condition, the solution is

$$\Delta V_{fg} = \frac{U_T}{\kappa} \ln\left(1 + \left(e^{\frac{\kappa \Delta V_{fg}(0^+)}{U_T}} - 1\right) e^{-\frac{t}{\tau_h}}\right). \tag{5.25}$$

After the initial jump, given by (5.19), the output voltage is related to the floating-gate voltage by

$$\Delta V_{out} = \frac{C_1 + C_2 + C_w}{C_2} \Delta V_{fg}. \tag{5.26}$$

Figure 5.10 shows measured AFGA output-voltage responses to several square-wave inputs. Figure 5.10a shows the responses of two unity-gain AFGAs with different capacitor values to the same square-wave input. As in the low-frequency case, the high-frequency response of the AFGA is asymmetric: the downgoing step response approaches its steady state linearly with time, and the upgoing step response approaches its steady state logarithmically with time. The initial jump in the downgoing step is due to capacitive feedthrough. From these data, it is evident that decreasing C_1 and C_2 without changing C_L will decrease the amount of capacitive feedthrough. Figure

5.10b shows the voltage responses, to a small input step, for two AFGAs with gains of 1 and 146. The response from the unity-gain AFGA is a buffered version of the input; the high-gain AFGA shows a linear, first-order, low-pass filtered version of the input. These responses illustrate the gain–bandwidth tradeoff in the AFGA.

The linear 3V output swing in the high-gain response of Fig. 5.10b raises this question: What determines the linear range of an AFGA? One criterion for linearity is that ΔV_{fg} be sufficiently small such that the factor $\left(\exp\left(-\frac{\kappa \Delta V_{fg}}{U_T}\right) - 1\right)$ in (5.20) can be approximated by $-\frac{\kappa \Delta V_{fg}}{U_T}$. This criterion implies that the floating-gate voltage must not move by more that $\frac{U_T}{\kappa}$ from its equilibrium value. The floating-gate voltage has its maximum swing in the capacitive-feedthrough regime; therefore, from (5.19), the input linear range, V_{Li}, is

$$V_{Li} = \frac{U_T}{\kappa}\left(\frac{C_1 + C_2 + C_w}{C_1}\right) B, \tag{5.27}$$

where I shall define

$$B = 1 - \frac{C_2^2}{(C_1 + C_2 + C_w)(C_L + C_2)}. \tag{5.28}$$

For amplifiers with gains greater than or equal to 1, which requires that C_1 be greater than C_2, B is bounded between $\frac{1}{2}$ and 1 for all C_1, C_2, C_W, and C_L. Further, if the AFGA is driving a C_L that is at least as big as C_1, B is bounded between $\frac{3}{4}$ and 1. Consequently, B can be considered a correction term.

I express the output linear range, V_{Lo}, in terms of the input linear range, V_{Li}, by

$$V_{Lo} = \frac{U_T}{\kappa}\left(\frac{C_1 + C_2 + C_w}{C_2}\right) B, \tag{5.29}$$

which is V_{Li} times the amplifier gain, C_1/C_2. The output linear range scales with the amplifier gain. By increasing C_w, I reduce the change in the floating-gate voltage, thereby increasing the amplifier's output linear range. The AFGA's gain from input to output in the passband is

$$\frac{V_{out}}{V_{in}} = -\left(\frac{C_1}{C_2}\right)\frac{1}{1 + \frac{C_1 + C_2 + C_W}{C_2 A}} = -\left(\frac{C_1}{C_2}\right)\frac{1}{1 + \frac{\kappa V_{Lo}}{A U_T B}}, \tag{5.30}$$

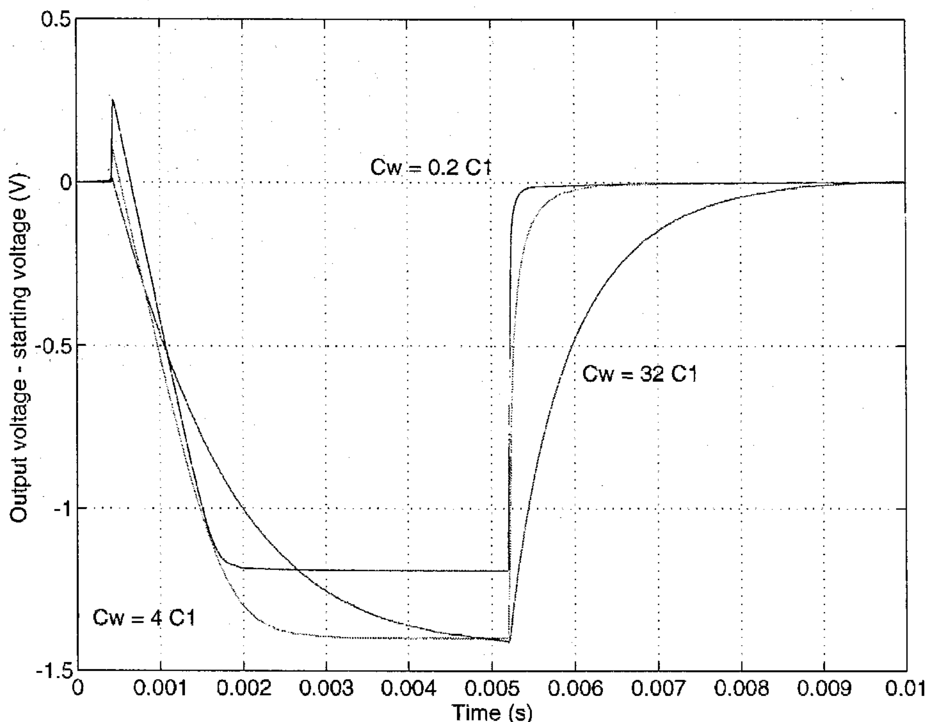


Figure 5.11: The response of three AFGAs to the same square-wave input. All three AFGAs were identical except for C_w, and were biased by the same V_τ. Increasing C_w increases the linear range, decreases the amount of capacitive feedthrough, and decreases the low-pass cutoff frequency.

where A is the gain from floating gate to output. For a sufficiently large A, the AFGA's passband gain is independent of C_w.

Figures 5.11 and 5.12 show measured data demonstrating how τ_h and linear range scale with C_w for unity-gain AFGAs. For a unity-gain AFGA—that is for $C_1 = C_2$—the expressions for τ_h and input linear range are

$$\tau_h = \frac{U_T(C_1 + C_L)}{\kappa I_\tau}\left(2 - \frac{C_1}{C_1 + C_L} + \frac{C_w}{C_1}\right), \tag{5.31}$$

and

$$V_{Li} = \frac{U_T}{\kappa}\left(2 - \frac{C_1}{C_1 + C_L} + \frac{C_w}{C_1}\right). \tag{5.32}$$

The data in Fig. 5.12 were taken with AFGAs that had no explicitly drawn C_L;

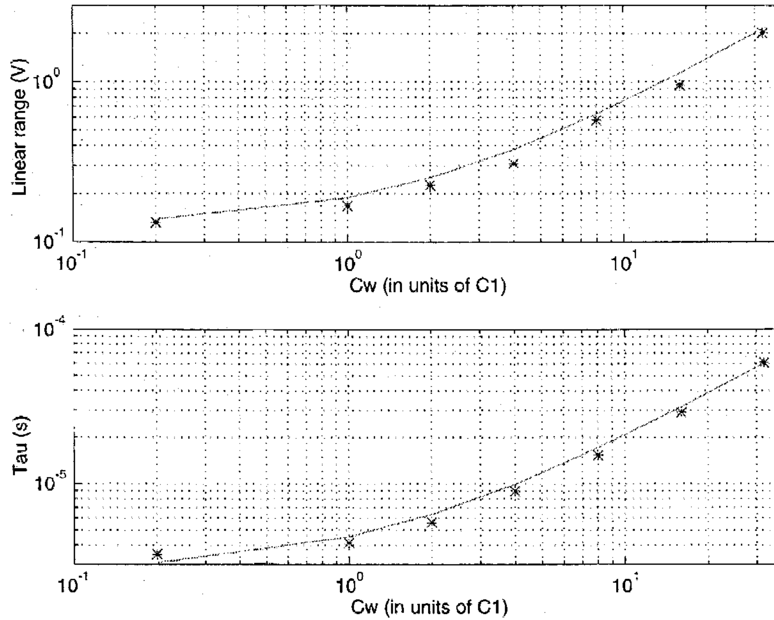


Figure 5.12: Measured linear range and τ_h for several unity-gain AFGAs for different C_w ratioed in units of C_1. The linear range fit is $V_{Li} = 0.063\text{V}\ C_w/C_1 + 0.125\text{V}$, and the τ fit is $\tau = 1.8\mu\text{s}\ C_w/C_1 + 2.7\mu\text{s}$.

the variation between the data and the linear curve fit is probably caused by the different parasitic load capacitances. Both the experimental data and the direct analytic solution of (5.22) indicate that second harmonic distortion dominates for the AFGAs; for a sine-wave input with amplitude of V_{Li}, the peak second harmonic distortion is 0.05 percent of (26dB below) the fundamental frequency response. The second harmonic distortion is maximum for frequencies just below $\frac{1}{2\pi\tau_h}$; for amplitudes at or below V_L, the second harmonic distortion is proportional to the square of the fundamental amplitude.

5.5 Frequency Response of the AFGA

To derive the AFGA frequency response, I begin with the small-signal form of

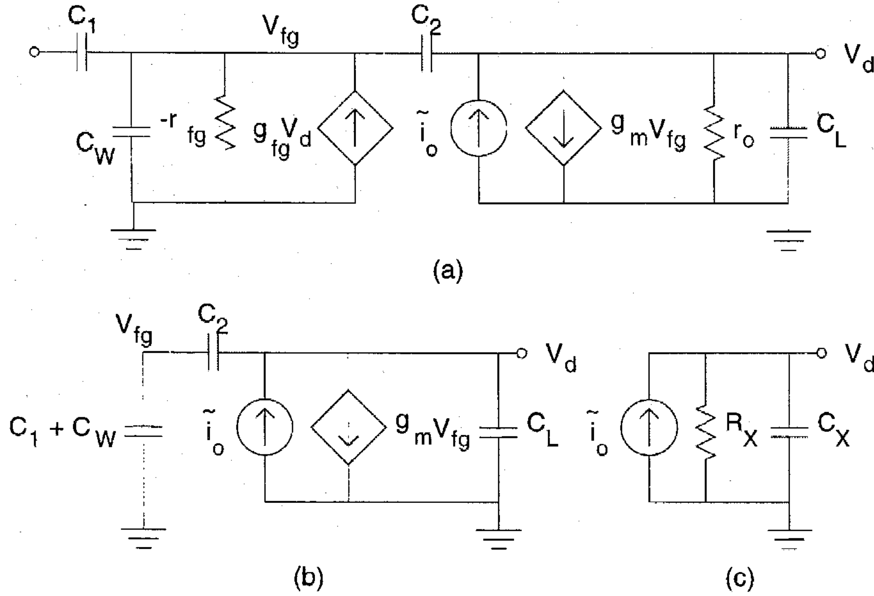


Figure 5.13: An AFGA represented as a small-signal circuit. (a) The small-signal AFGA model using the small-signal pFET model. (b) The small-signal model of the effect of the noise source in the channel on the output voltage. I have neglected the effect of the gate current, as well as the Early voltage effect, in this model. (c) A simplified small-signal model of the effect of noise. For clarity, I define $R_x = \frac{C_1+C_2+C_w}{g_m C_2}$, and $C_x = C_L + C_2\left(1 - \frac{C_2}{C_1+C_2-C_w}\right)$

(5.7) and (5.8):

$$(C_1 + C_2 + C_w)\frac{dV_{fg}}{dt} = C_1\frac{dV_{in}}{dt} + C_2\frac{dV_{out}}{dt} + \frac{I_{tun0}}{V_{inj}}\Delta V_{out},$$

$$(C_2 + C_L)\frac{dV_{out}}{dt} = C_2\frac{dV_{fg}}{dt} - \frac{\kappa I_T}{U_T}\Delta V_{fg}; \tag{5.33}$$

that is, I assume that the input signal is sufficiently small that I need to keep only the linear terms when we expand the exponentials. A small-signal input changes V_{out} by less than V_{inj}, due to the injection nonlinearity in the low-frequency regime, and change V_{fg} by less than $\frac{U_T}{\kappa}$, due to the transistor nonlinearity in the high-frequency regime. We can also obtain (5.33) by analyzing the small-signal circuit in Fig. 5.13a. I first discuss the response in the low- and high-frequency regimes, and then present the general solution.

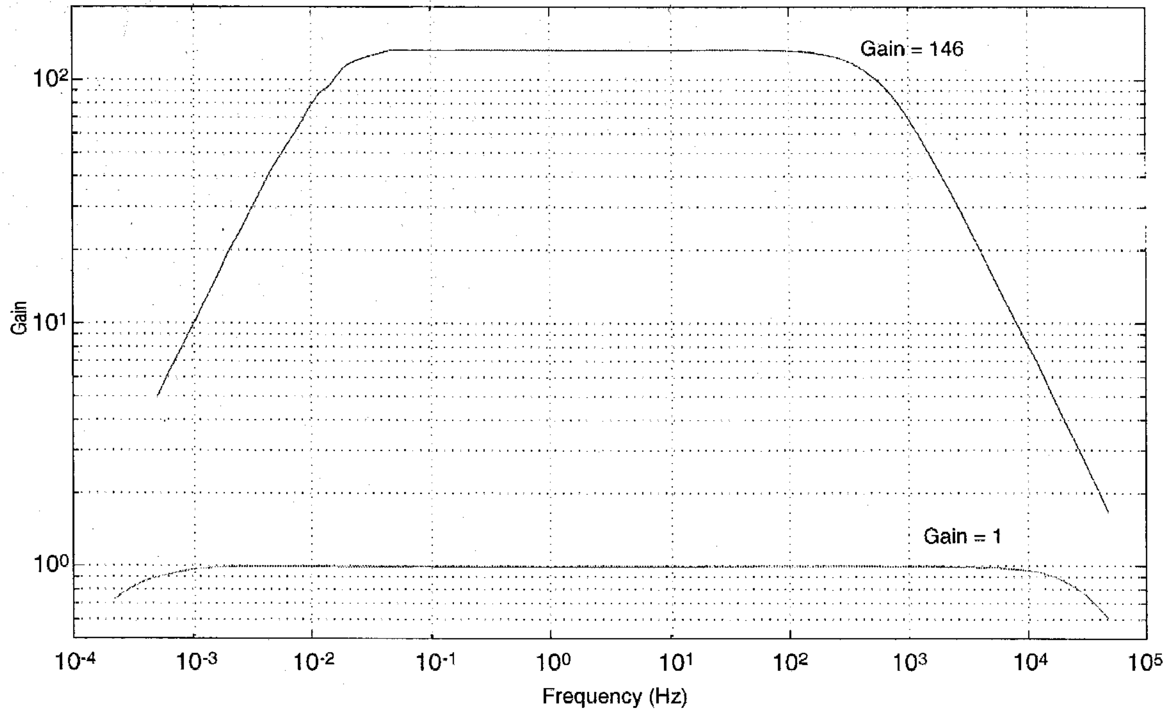


Figure 5.14: Frequency response for two AFGAs with different gains. For both the high- and low-gain AFGA, $C_1 + C_2$ is approximately constant. For the high-gain AFGA, τ_l is 20mHz, and τ_h is 600Hz; for the low-gain AFGA, τ_l is 300μHz and τ_h is 40kHz. The ratio of τ_h and τ_l between the two AFGAs are equal to one-half of the ratio of the gains; the ratio is consistent with a constant $C_1 + C_2$.

For low-frequency inputs, I approximate (5.33) as

$$\tau_l \frac{C_1}{C_2}\frac{dV_{in}}{dt} + \tau_l \frac{dV_{out}}{dt} = -\Delta V_{out}, \tag{5.34}$$

for which the resulting frequency response is

$$\frac{V_{out}(s)}{V_{in}(s)} = -\frac{C_1}{C_2}\frac{s\tau_l}{1 + s\tau_l}. \tag{5.35}$$

Figure 5.14 shows the measured AFGA frequency response: for the high-gain AFGA, τ_l is 20mHz; for the low-gain AFGA, τ_l is 300μHz. The high-gain AFGA has a gain of 146; the low-gain AFGA has unity gain.

For high-frequency inputs, I can simplify (5.33) by assuming input frequencies

much larger than $\frac{1}{2\pi\tau_l}$; I write the result as

$$\frac{V_{out}}{V_{in}} = -\frac{C_1}{C_2}\frac{1 - \tau_{h2} s}{1 + \tau_h s}. \tag{5.36}$$

This transfer function includes the effects of parasitic and load capacitances. The response in (5.36) is the transfer function of a first-order system; because we use capacitive feedback, the AFGA is stable for any value of closed-loop gain. As we can see in Fig. 5.14, $\frac{1}{2\pi\tau_h}$ is 500Hz for the high-gain AFGA, and is 40kHz for the low-gain AFGA.

To obtain the response for all frequencies, we can take the Laplace transform of (5.33):

$$s\left(C_1 + C_2 + C_w\right) V_{fg} = sC_1 V_{in} + \left(sC_2 + \frac{I_{tun0}}{V_{inj}}\right) V_{out},$$

$$s\left(C_2 + C_L\right) V_{out}(s) = \left(sC_2 + \frac{\kappa I_\tau}{U_T}\right) V_{fg}. \tag{5.37}$$

We can solve (5.37) to obtain

$$\frac{V_{out}(s)}{V_{in}(s)} = -\frac{C_1}{C_2}\frac{1 - \tau_{h2} s}{1 + \tau_h s + \frac{1}{\tau_l s}}, \tag{5.38}$$

where $\tau_l, \tau_h, \tau_{h2}$ are as defined previously.

When considering the frequency response of the AFGA, it is natural to consider the output-voltage spectrum for no input—that is, the output-voltage noise from the amplifier. Figures 5.15 and 5.16 show AFGA output-voltage spectra for a fixed, voltage-source input. For low frequencies, $1/f$ noise is dominant; for high frequencies, thermal noise dominates. The AFGA attenuates the $1/f$ noise below the low-frequency cutoff. Figure 5.15 shows that we can reduce the $1/f$ noise by increasing V_{tun}, and thereby decreasing τ_l. Figure 5.16 shows a comparison among a high-gain AFGA, a unity-gain AFGA, and a follower-connected transconductance amplifier. The transconductance amplifier is the wide-range amplifier described previously [82]; it has transistors larger than those of the AFGAs, resulting in the lower $1/f$ noise. The AFGAs used a constant tunneling current; because the noise spectrum of the unity-

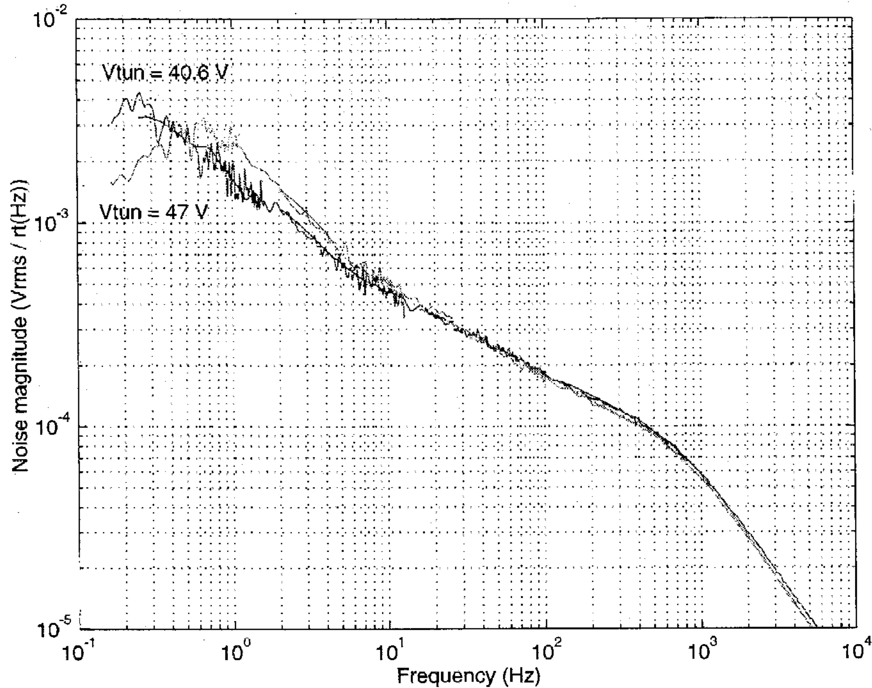


Figure 5.15: Output noise spectrum of an AFGA with a gain of 146 for two different tunneling voltages (V_{tun}) and a constant input. The high-frequency cutoff eliminates $1/f$ noise at frequencies below 1 / $2\pi\tau_l$. The spectrum was taken for a bias current of 80nA, which corresponds to a V_T of 0.73V.

gain AFGA is not appreciably different from that of the transconductance amplifier, we conclude that the tunneling and injection processes do not contribute significantly to the noise levels.

Let us investigate how changing the AFGA design will change the amount of output noise. Following [88], we can model the thermal noise component, $\hat{i}_o$, of a subthreshold MOSFET's channel current by

$$\frac{\hat{i}_o^2}{\Delta f} = \frac{2}{\kappa} q U_T g_m. \tag{5.39}$$

Because the AFGA's output comprises both an nFET and a pFET, the total thermal-noise current derives from two parallel noise sources. I want to find the output-referred voltage noise, which I obtain from the simplified small-signal circuit of Fig. 5.13(b). I can further simplify the small-signal circuit of Fig. 5.13(b) to that

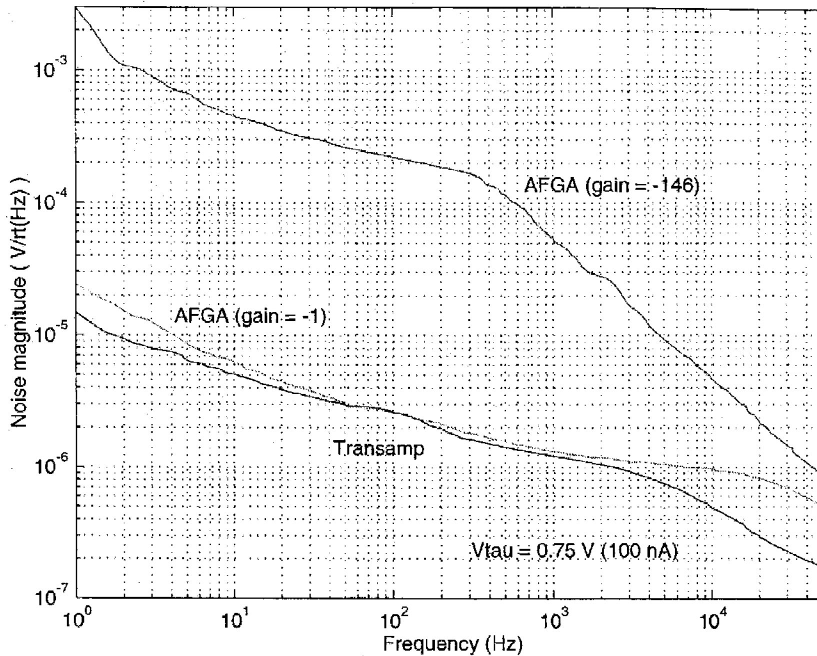


Figure 5.16: Comparison of a high-gain AFGA with a unity-gain AFGA and with a generic follower-connected differential amplifier. All three amplifiers had the same V_τ voltage, and had the same bias current. The sums of C_1 and C_2 are the same for the two AFGAs.

shown in Fig. 5.13(c), by noting that I can relate V_{fg} to V_{out} by a capacitive divider. From this simpler circuit, we can express the signal power of the output-referred voltage noise, $\hat{V}_{out}^2$, as

$$\hat{V}_{out}^2 = \left(\frac{C_1 + C_2 + C_w}{C_2 g_m}\right)^2 \frac{\hat{i}_o^2}{1 + (\omega\tau_h)^2}, \tag{5.40}$$

where τ_h is as defined in Section 5.4. From this expression, we can calculate the total output-noise power as

$$\hat{V}_{out}^2 = \frac{4}{\kappa} q U_T g_m \left(\frac{C_1 + C_2 + C_w}{C_2 g_m}\right)^2 \int_0^\infty \frac{1}{1 + (\omega\tau_h)^2} df, \tag{5.41}$$

which, when we use (5.24), evaluates to

$$\hat{V}_{out}^2 = \frac{qU_T}{\kappa B}\frac{C_1 + C_2 + C_w}{C_2(C_L + C_2)}, \tag{5.42}$$

where the correction term, B, is as defined in (5.28). The total output-noise power is roughly proportional to C_w, and is inversely proportional to C_L.

Now, I would like to calculate the AFGA dynamic range. I shall define dynamic range, DR, as the ratio of the maximum possible linear output swing to the total output-noise power. With this definition, which is equivalent to that given in[89], I express the AFGA dynamic range as

$$DR = \frac{V_{Lo}^2}{2\hat{V}_{out}^2} = \frac{\kappa}{2q} V_{Lo}(C_L + C_2)B^2, \tag{5.43}$$

which is similar to the form for dynamic range for the wide-linear-range amplifier, as derived in [89]. The dynamic range varies inversely with C_2; therefore a high-gain amplifier will have a larger dynamic range than will the low-gain amplifier for the same values of C_1, C_w, and C_L.

5.6 Steady-State Output-Voltage Dependence on the Output Signal

This section discusses the dependence of the steady-state output voltage on the input signal amplitude for input frequencies in the AFGA's passband. Figure 5.17 shows measurements of the minimum and maximum output voltages versus the output amplitude. For small input amplitudes, the minimum and maximum output voltages deviate symmetrically from the steady-state voltage, but for large input amplitudes, most of the change in the output voltage is due to an increasing maximum output voltage. In general, the steady-state output voltage remains within about V_{inj} of the minimum of the signal.

To analyze this effect, I decompose the output voltage into components that

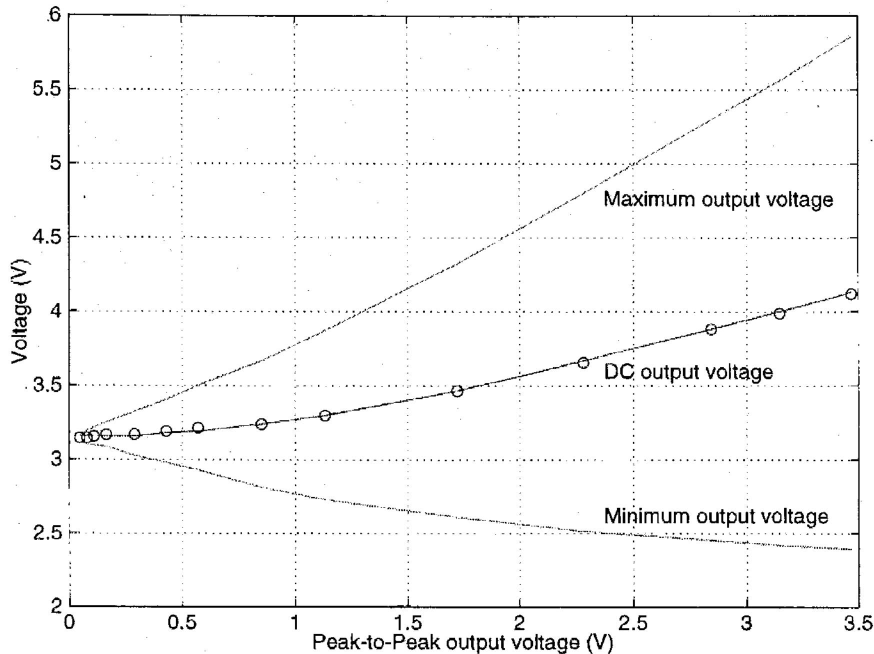


Figure 5.17: Minimum and maximum output voltages versus the peak-to-peak output-voltage amplitude. The frequency of the input sine wave was 100Hz; the AFGA had a gain of 146. For small input amplitudes, the minimum and maximum output voltages symmetrically deviate from the steady-state voltage; for large input amplitudes, however, the DC output voltage follows the maximum output voltage. The DC voltage was fit to the function $0.5\ln(\mathrm{I}_0(V_{dc}\ /\ 1.0\mathrm{V}\ )\ )$, which is equal to (5.53) with $V_{inj} = 500\mathrm{mV}$.

change at fast and slow rates. As in Chapter 3, I assume that we can express the output voltage as a sum of variables that represent the fast and slow rates; that is,

$$\Delta V_{out} = \Delta \hat{V}_{out} + \Delta \bar{V}_{out}, \tag{5.44}$$

where $\Delta \hat{V}_{out}$ represents the fast-timescale behavior, and is the amplified version of ΔV_{in} ($\Delta \hat{V}_{out} \approx -\frac{C_1}{C_2}\Delta V_{in}$), and $\Delta \bar{V}_{out}$ represents the slow-timescale behavior. With this formulation, we can integrate (5.44) over many periods at the fast timescale, but still make only a small change in the slow-timescale output voltage. I shall define $E[\cdot]$ as the average of a time-varying signal, $x(t)$, over a time interval, T, that is much shorter than the slow timescale, but much longer than one period at the fast

timescale:

$$E[x(t)] = \frac{1}{T}\int_0^T x(t)dt. \tag{5.45}$$

By this definition,

$$E\left[\frac{dV_{in}}{dt}\right] \to 0, E\left[\frac{d\bar{V}_{out}}{dt}\right] \to \frac{d\bar{V}_{out}}{dt}. \tag{5.46}$$

The resulting equation in $\Delta\bar{V}_{out}$ from (5.11) is

$$C_2\frac{d\bar{V}_{out}}{dt} = I_{tun0}\left(E\left[e^{-\frac{\Delta V_{out}}{V_{inj}}}\right] - 1\right). \tag{5.47}$$

We need to express $E\left[e^{-\frac{\Delta V_{out}}{V_{inj}}}\right]$ in terms of the fast and slow variables:

$$E\left[e^{-\frac{\Delta V_{out}}{V_{inj}}}\right] = e^{-\frac{\Delta\bar{V}_{out}}{V_{inj}}}Q, \tag{5.48}$$

where

$$Q = E\left[e^{-\frac{\Delta\hat{V}_{out}}{V_{inj}}}\right]. \tag{5.49}$$

I rewrite (5.47) as

$$C_2\frac{d\bar{V}_{out}}{dt} = I_{tun0}\left(Qe^{-\frac{\Delta\bar{V}_{out}}{V_{inj}}} - 1\right). \tag{5.50}$$

From the analysis in Section 5.3, the solution to (5.50) is

$$\Delta\bar{V}_{out} = V_{inj}\ln\left(Q + (e^{\frac{\bar{V}_{out}(0^-)}{V_{inj}}} - Q)e^{-\frac{t}{\tau}}\right), \tag{5.51}$$

where the steady-state solution for $\Delta\bar{V}_{out}$ is

$$\Delta\bar{V}_{out} = V_{inj}\ln\left(E\left[e^{-\frac{\Delta\hat{V}_{out}}{V_{inj}}}\right]\right). \tag{5.52}$$

The AFGA always adapts its floating-gate charge such that the minimum of the output signal remains at the equilibrium output voltage.

Now, let us consider the output-voltage behavior as a function of the input-signal amplitude, for a sinusoidal input. I define the amplified input signal as $\Delta\hat{V}_{out} =$

$A \sin(\omega t)$; for this output signal, the steady-state voltage is

$$\Delta\bar{V}_{out} = V_{inj} \ln \left(\int_{\omega t=0}^{2\pi} e^{-\frac{A}{V_{inj}} \sin(\omega t)} d(\omega t) \right) = V_{inj} \ln \left(\mathrm{I}_0 \left(\frac{A}{V_{inj}} \right) \right), \tag{5.53}$$

where $\mathrm{I}_0(\cdot)$ is the modified Bessal function of zero_{th} order. Figure 5.17 shows measured minimum and maximum output voltages versus the output signal amplitude for a sine-wave input.

5.7 Other AFGA Effects

5.7.1 Model of the Above-Threshold AFGA

The operation of an autozeroing floating-gate amplifier with above threshold bias currents is similar to the subthreshold behavior in three important respects. First, those effects that depend on electron tunneling remain the same, because tunneling is not a function of the MOSFET channel current. Second, the injection current is still the exponential of the drain voltage. Third, the low-frequency dynamics remain unchanged for a constant C_2, because the channel current is held fixed by feedback.

The pFET hot-electron injection model changes for above-threshold bias currents in two ways. First, the source current is no longer an exponential function of the gate voltage, but rather varies quadratically with the gate voltage. Second, the injection current decreases relative to the source current, because the impact-ionization efficiency will decrease as a result of the potential drop along the channel. I modify the above-threshold injection-current model from (3.17) to be

$$\begin{aligned} I_{inj} &= I_{inj0} \left(\frac{V_{dd} - V_{fg} - \Delta V_{fg} + V_T}{V_{dd} - V_{fg} + V_T} \right)^2 e^{-\frac{\Delta V_d - \kappa \Delta V_g}{V_{inj}}} \\ &\approx I_{inj0} \left(1 - \frac{2\Delta V_{fg}}{V_{dd} - V_{fg} + V_T} \right) e^{-\frac{\Delta V_d - \kappa \Delta V_g}{V_{inj}}} \\ &\approx I_{inj0} \exp \left(-\frac{2\Delta V_{fg}}{V_{dd} - V_{fg} + V_T} \right) e^{-\frac{\Delta V_d}{V_{inj}}}. \end{aligned} \tag{5.54}$$

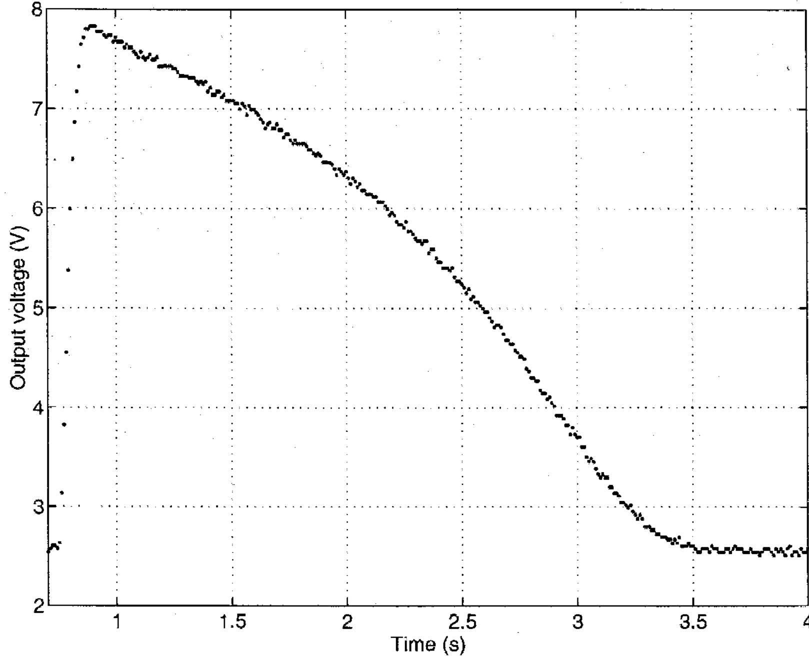


Figure 5.18: The response of an above-threshold AFGA to a downgoing step The feedback cap, C_2, of this AFGA is only the parasitic floating-gate–to–drain overlap capacitance. In the subthreshold case, the voltage linearly decreases with time; therefore, the nonlinear decrease of the output voltage for an above-threshold bias shows that the overlap capacitance changes with drain voltage.

This model changes the DC output voltage to

$$\Delta V_d = -\frac{V_{inj}}{V_x}\Delta V_{tun} + V_{inj}\frac{\kappa_n}{\kappa}\left(\frac{2}{V_{dd} - V_{fg} + V_T} - \frac{1}{V_x}\right)\Delta V_\tau. \tag{5.55}$$

Figure 5.5 shows data consistent with (5.55): the equilibrium output voltage saturates for above-threshold bias currents.

Now consider how the high-frequency behavior changes for above-threshold bias currents. For above-threshold currents, I modify (5.20) to

$$\tau_h \frac{dV_{fg}}{dt} = \tau_h 2\frac{dV_{in}}{dt} - \Delta V_{fg}\left(1 - \frac{\kappa \Delta V_{fg}}{V_{dd} - \kappa(V_{fg0} + V_{T0})}\right). \tag{5.56}$$

The above threshold definitions of τ_h and τ_{h2} are:

$$\tau_{h2} = \frac{C_1(C_2 + C_L)(V_{dd} - \kappa(V_{fg0} + V_{T0}))}{\kappa C_2 I_\tau}, \tag{5.57}$$

and

$$\tau_h = \left((C_1 + C_2 + C_w)(C_2 + C_L) - C_2^2\right) \frac{V_{dd} - \kappa(V_{fg0} + V_{T0})}{\kappa C_2 I_\tau}. \tag{5.58}$$

The higher bias currents result in a higher cutoff frequency, but also requires an increase in power dissipation. The linear input range also is larger, and now is $\frac{C_1+C_2+C_w}{\kappa C_1}\left(V_{dd} - \kappa(V_{fg0} + V_{T0})\right)$. The large-signal dynamics change for above-threshold biases. The floating-gate–to–drain overlap capacitance becomes a function of the drain voltage, adding additional dynamics to the large-signal response. Figure 5.18 shows the response of an above-threshold AFGA to a downgoing step.

5.7.2 Continuous Operation of the Tunneling Current

At this point, we might ask what happens when we do not use a continuous tunneling current in the AFGA. The constant tunneling bias current naturally eliminates the effect of DC biasing points; without the gate current, we no longer have an autozeroing amplifier. Without the autozeroing behavior, we need to add circuitry to remove the DC offset. Furthermore, there is additional $1/f$ noise, because the AFGAs high-pass behavior filters the low-frequency $1/f$ noise.

Suppose that we autozero the floating-gate amplifier only to set a particular operating point. Then, after this calibration phase, we lower the tunneling voltage and power supply to turn off the electron-tunneling and hot-electron injection processes to eliminate the gate currents. Unfortunately, the capacitive coupling to the floating gate by the tunneling junction and capacitances not referenced to V_{dd} will make potentially large changes in the output voltage. We can minimize this effect by using small tunneling junctions and by ensuring that all the capacitances (including all parasitics) that couple to the floating gate are referenced to V_{dd}.

In addition to the capacitive coupling, the charge on the floating gate will change

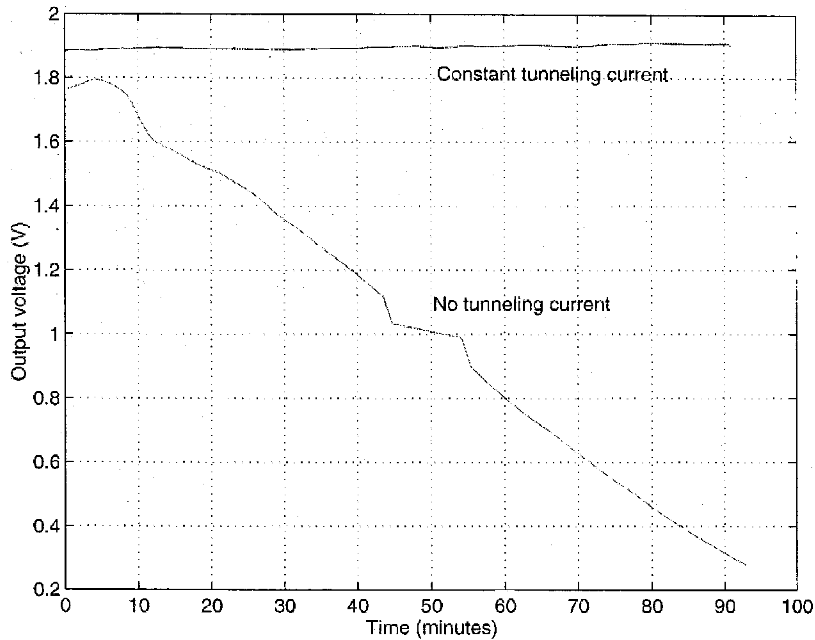


Figure 5.19: Change in the AFGA output voltage with and without a continuous tunneling current. I used the same AFGA with a gain of 146 with both experiments. The experiment for no tunneling current also required that I drop the power supply; V_{dd} was set at 5V. The trace with no tunneling current was started 5 minutes after the tunneling line was dropped.

due to electron traps in the MOS gate oxides. When we tunnel or inject electrons into SiO_2, a portion of these electrons are trapped in the oxide [90]. In addition, as more current passes through these oxides, more electron traps are created [91]. Once the oxide currents stop, some of these trapped electrons will detrap, and can find their way to the floating gate. This effect will result in a large drift in the output voltage over time for the same input bias voltage. In Fig. 5.19, I show the change in the output voltage over time for a high-gain AFGA with and without a constant tunneling current. The detrapping shows no sign of stopping until the output runs into ground; in a lower-gain device, the detrapping may eventually settle with the output voltage within the supply rails.

5.7.3 Restoration to Equilibrium when the Output Voltage Starts at a Supply Rail

Next, we might ask what happens when the output voltage starts at one of the supply rails. The output voltage starts at V_{dd} when the floating-gate voltage is too low. In this regime, there is no injection current; the tunneling current removes electrons from the floating gate, raising the floating-gate voltage. Eventually, the floating-gate voltage increases to its steady-state level, and the output voltage decreases from V_{dd}. If, on the other hand, the output voltage starts near ground, then the pFET has insufficient channel current to develop enough injection to balance the tunneling current. The current must be balanced by changes in the floating-gate voltage: In this case, the floating-gate voltage will increase, since the tunneling current exceeds the pFET injection current. Unfortunately, the pFET channel current then decreases even further, in turn decreasing the injection current, and leading to a runaway condition. The steady-state output voltage may not return to the original equilibrium level. Typically, this condition poses no problem even for reasonably large changes in the input voltage; however, we see this effect at startup or when we make large changes in V_{τ}. Decreasing V_{tun}, decreasing the input voltage, increasing V_{τ}, or increasing V_{dd} might allow the AFGA to recover from this condition.

5.8 Conclusions

The AFGA is a simple example of a large class of adaptive floating-gate MOS circuits; these circuits use tunneling and hot-electron injection to adapt the charge on floating gates to return the circuit to a baseline condition on a slow timescale. When the appropriate feedback is applied to the floating gate, this adaptation is an inherent part of the circuit's operation—no additional control circuitry is required. In the case of AFGA, I set up the feedback such that the output voltage returns to its steady-state value on a long timescale. The modulation of the pFET hot-electron injection by the output voltage provides the correct feedback to return the output voltage to

the proper operating regime.

The AFGA has four operating regimes. First, in the adaptation regime, the AFGA behaves as a high-pass filter; the timescale is set by the tunneling and injection currents. Second, in the integrating regime, the AFGA behaves as a low-pass filter; the timescale is set by the nFET bias current. Third, for timescales between the adaptation and integrating regimes, the AFGA acts as an amplifier. Fourth, at frequencies much higher than the integrating regime, the AFGA exhibits capacitive feedthrough, which can be reduced by an increase in either C_w or C_L.

The AFGA always is a first-order system, even in the presence of parasitic capacitances; therefore, the AFGA is unconditionally stable, with 90 degrees of phase margin for noninductive loads. An amplifier that has resistive feedback is at least a second-order system, but an amplifier with capacitive feedback can be a first-order system.

MOS devices and quantum processes, such as electron tunneling and hot-electron injection, are often criticized for their high $1/f$ noise. Since the AFGA's noise performance is similar in thermal and $1/f$ characteristics to that of a standard MOS amplifier, the tunneling and injection processes do not add appreciable noise to the amplifier. In addition, with a desired adaptation rate, I can reduce significantly the low-frequency noise generated in the AFGA; such a reduction cannot be obtained in a standard amplifier that has a blocking capacitor at the input. For moderate tunneling currents, the low-frequency time constant can remain nearly constant for timescales measured in years; any shift is due primarily to trapping in the tunneling oxide. I can increase the linear range by increasing C_w, and I can increase the dynamic range by increasing C_w or C_L.

Chapter 6 Conclusions and Future Directions

This dissertation described the foundations of a floating-gate technology for adaptation and learning. I developed an analytical model of hot-electron injection and impact ionization from first principles, and described supportive experimental measurements from subthreshold nFETs and pFETs. I described the invention, characterization, and modeling of the single-transistor learning synapses. Then, I characterized and modeled the continuous-time floating-gate dynamics of simple circuits of single-transistor learning synapses. Finally, I described the invention, characterization, and modeling of the autozeroing floating-gate amplifier. These four building blocks form the foundation for several large-scale floating-gate adaptive systems.

6.1 Summary of Thesis Results

Chapter 2 presented a model hot-electron transport in the drain-to-channel depletion region using the spatially varying Boltzmann transport equation, and analytically found a self-consistent distribution function in a two-step process. In the first step, I solved for the average hot-electron trajectory in energy and direction as a function of position through the depletion region. In the second step, I solved for the electron distribution function around this average electron trajectory. In this coordinate system, phonon collisions spatially diffuse the distribution function, and impact-ionization collisions remove electrons from the high-energy distribution function. From the electron distribution function, I calculated the probabilities of impact ionization and hot-electron injection as functions of channel current, drain voltage, and floating-gate voltage. I compared the results of my analytical model to measurements in long-channel devices. The model simultaneously fits both the hot-electron-injection and

impact-ionization data. These analytical results yield an energy-dependent impact-ionization collision rate that is consistent with numerically calculated collision rates reported in the literature.

To make the transition from devices to circuits, Chapter 3 presented the single-transistor learning synapses. The single-transistor synapse, invented and developed by me and my collaborators, Diorio and Minch, is an hot-electron injecting FET with a well-tunneling junction. The single-transistor synapses simultaneously perform long-term weight storage, compute the product of the input and the weight value, and update the weight value according to a Hebbian or a backpropagation learning rule. Memory is accomplished via charge storage on polysilicon floating gates, providing long-term retention without refresh. The small size and low power operation of single-transistor synapses allows the development of dense synaptic arrays. The synapses efficiently use the physics of silicon to perform weight updates; electron tunneling increases the weight value and hot-electron injection decreases the weight value. I presented a model of electron tunneling, which I use to remove electrons from the floating gate, and I presented hot-electron-injection measurements in both nFET and pFETs. The charge on the floating gate is decreased by hot-electron injection with high selectivity for a particular synapse. The charge on the floating gate is increased by electron tunneling, which results in high selectivity between rows, but in much lower selectivity between columns along a row.

Chapter 4 presented the possible negative- and positive-feedback configurations of continuous-time floating- gate MOS circuits. The learning synapses provide different floating-gate dynamics depending on their configuration. I built a new type of pFET synapse by degenerating the source; the oxide currents of the synapse provide stabilizing feedback to the floating gate and to the drain. I discussed the range of possible stabilizing and destabilizing types of feedback in circuits with one floating-gate synapse, including data from nFET, pFET, and source-degenerated pFET synapses. Multiple floating-gate circuits show competitive and cooperative behavior between synapses, and that is consistent with Hebbian and anti-Hebbian learning. Modeling the dynamics of a floating-gate circuit builds intuition to develop more complex

circuits, and shows the strengths of simple floating-gate current models.

The autozeroing floating-gate amplifier (AFGA) is a simple example of this class of adaptive floating-gate MOS circuits; these circuits use tunneling and hot-electron injection to adapt the charge on floating gates to return the circuit to a baseline condition on a slow timescale. When the appropriate feedback is applied to the floating gate, this adaptation is an inherent part of the circuit's operation—no additional control circuitry is required. The modulation of the *p*FET hot-electron injection by the output voltage provides the correct feedback to return the output voltage to the proper operating regime.

Chapter 5 presented a bandpass floating-gate amplifier that uses tunneling and *p*FET hot-electron injection to set its DC operating point adaptively. Because the hot-electron injection is an inherent part of the *p*FETs behavior, I obtain this adaptation with no additional circuitry. Because the gate currents are small, the circuit exhibits a high-pass characteristic with a cutoff frequency less than 1 Hz. The high-frequency cutoff is controlled electronically, as is done in continuous-time filters. I have derived analytical models that completely characterize the amplifier and that are in good agreement with experimental data for a wide range of operating conditions and input waveforms. The AFGA always is a first-order system, even in the presence of parasitic capacitances; therefore, the AFGA is unconditionally stable, with 90 degrees of phase margin for noninductive loads. Since the AFGA's noise performance is similar in thermal and $1/f$ characteristics to that of a standard MOS amplifier, the tunneling and injection processes do not add appreciable noise to the amplifier. In addition, with a desired adaptation rate, I can reduce significantly the low-frequency noise generated in the AFGA. I can increase the linear range by increasing C_w, and I can increase the dynamic range by increasing C_w or C_L. This autozeroing floating-gate amplifier demonstrates how to use continuous-time, floating-gate adaptation in amplifier design.

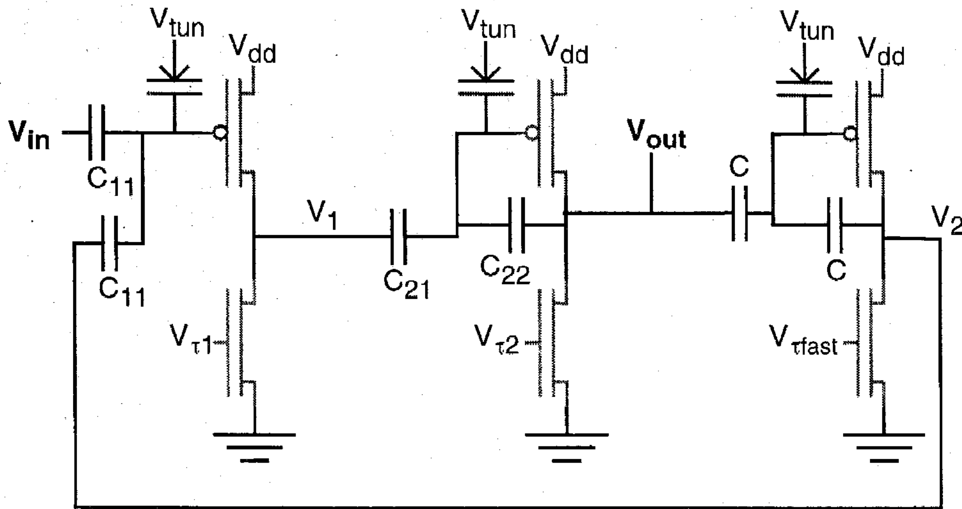


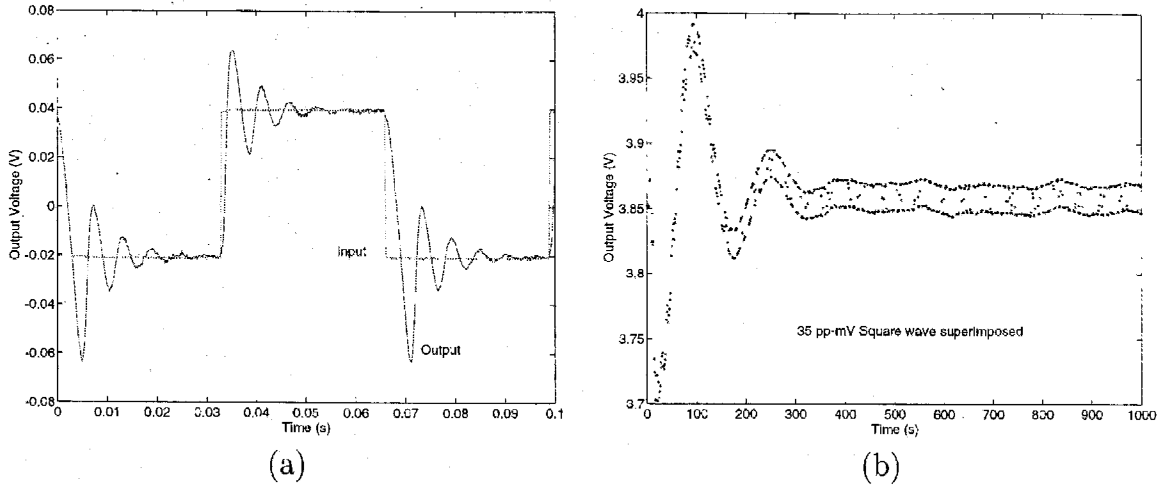
Figure 6.1: Circuit diagram of the autozeroing second-order section. This circuit, which is built with three autozeroing amplifiers, shows second-order behavior that is electronically controlled.

6.2 Future Directions

This thesis considered the spectrum of a silicon floating-gate technology for adaptation and learning spanning from hot-electron transport to simple floating-gate circuits and dynamics. The remaining question is whether this technology will scale to large floating-gate circuits and systems with adaptation and learning. The rest of this section briefly presents two floating-gate circuits, the autozeroing second-order section and the adaptive winner-take-all; these circuits show that this technology can scale to larger circuits and systems.

6.2.1 Autozeroing Second-Order Section

Figure 6.1 shows my autozeroing second-order section; it is the primary element to build higher-order filters, both in continuous-time and switch-capacitor circuits. This circuit is based on the Diff2 second-order section [82], which in turn is based on a canonical form for a second-order section as a high-gain amplifier with feedback. To build a simplified model of the second-order behavior, we assume that $V_2 = -V_{out}$; that is, $V_{\tau fast}$ is set at a bias current much larger than $V_{\tau 1}$ or $V_{\tau 2}$. We ignore (for



(a) (b)

Figure 6.2: Step response of the autozeroing second-order section. (a) Short timescale relaxation to upgoing and downgoing input steps. The ringing of the output voltage is characteristic of a second-order system. (b) Long timescale relaxation to an upgoing input steps; a 1Hz square wave was superimposed on the input signal, and is preserved throughout the relaxation. This ringing behavior proves that the circuit exhibits at least second-order behavior from the AFGA corner frequencies set by the floating-gate currents.

the moment) the floating-gate currents, ignore the capacitive feedthrough effects, and assume the $V_{\tau 1}$ amplifier's gain is very large. From the AFGA analysis in Chapter 5, the small-signal model of the two amplifiers is

$$\frac{C_{T1}C_{o1} - C_{21}^2}{C_{21}} \frac{U_T}{\kappa I_{\tau 1}} \frac{dV_1}{dt} = (V_{in} + V_2), \tag{6.1}$$

$$\frac{C_{T2}C_{o2} - C_{22}^2}{C_{22}} \frac{U_T}{\kappa I_{\tau 2}} \frac{dV_{out}}{dt} = (V_1 + V_{out}), \tag{6.2}$$

where C_T is the total capacitance connected to the floating gate ($C_T = C_1 + C_2 + C_w$), C_o is the total capacitance connected to the output node ($C_o = C_2 + C_L$), and C_{21}, C_{22} are the feedback capacitors from floating gate to output. By taking the Laplace transform of these equations, we solve for the transfer function from V_{in} to V_{out} as

$$\frac{V_{out}(s)}{V_{in}(s)} = \frac{1}{1 + s\frac{C_{T1}C_{o1} - C_{21}^2}{C_{21}} \frac{U_T}{\kappa I_{\tau 1}} + s^2 \frac{C_{T1}C_{o1} - C_{21}^2}{C_{21}} \frac{CU_T}{\kappa I_{\tau 1}} \frac{C_{T2}C_{o2} - C_{22}^2}{C_{22}} \frac{CU_T}{\kappa I_{\tau 2}}} = \frac{1}{1 + s\tau_1 + s^2 \tau_1 \tau_2}, \tag{6.3}$$

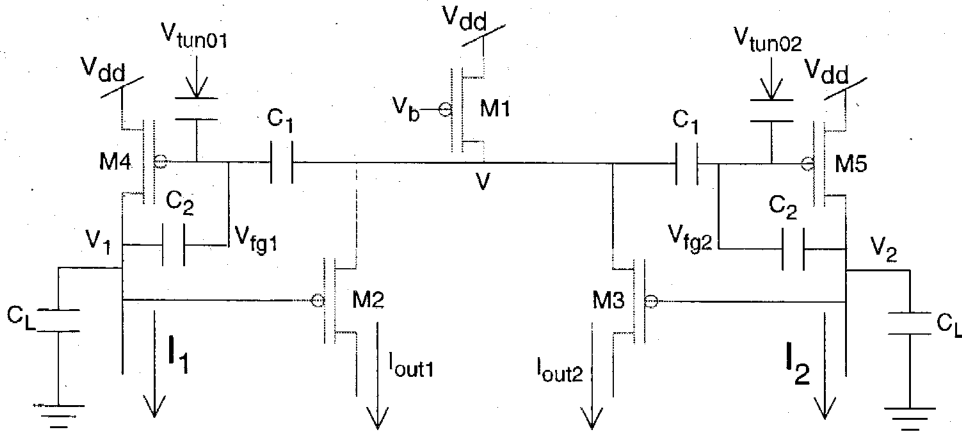


Figure 6.3: The circuit diagram of a two-input winner-take-all circuit.

where I define

$$\tau_1 = \frac{C_{T1}C_{o1} - C_{21}^2}{C_{21}} \frac{U_T}{\kappa I_{\tau 1}}, \text{and } \tau_2 = \frac{C_{T2}C_{o2} - C_{22}^2}{C_{22}} \frac{U_T}{\kappa I_{\tau 2}}. \tag{6.4}$$

From the canonical form for the second-order section, we find that [82]

$$\tau = \sqrt{\tau_1 \tau_2}, \text{and } Q = \sqrt{\frac{\tau_1}{\tau_2}}. \tag{6.5}$$

Figure 6.2(a) shows the step response of this autozeroing second-order section. The data show the characteristic ringing behavior of a second-order system; changing $V_{\tau 1}$ and $V_{\tau 2}$ changes τ and Q, as predicted from (6.5). Figure 6.2(b) shows the step response at a much slower timescale. The oscillatory behavior is due to the corner frequencies, which are set by the floating-gate currents. This part of the dynamics is usually third order, and therefore the modeling becomes more difficult.

6.2.2 Adaptive Winner-Take-All

Figure 6.3 shows the two-input, adaptive winner–take- all (WTA) circuit, invented by W. Fritz Kruger and myself [93], based on Lazzaro's classic WTA [92]. We have added a time dimension (adaptation) to make the input derivative an important factor in

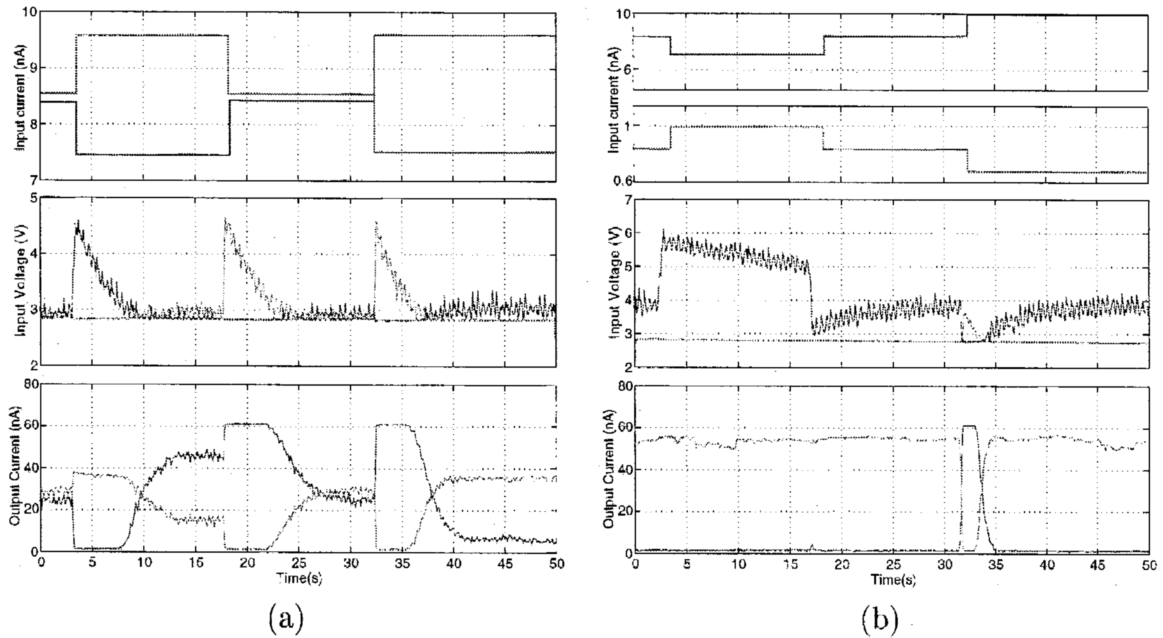


Figure 6.4: Time traces of the output current and voltage for small differential input current steps. (a) Time traces for small differential current steps around nearly identical bias currents of 8.6nA. (b) Time traces for small differential current steps around two different bias currents of 8.7nA and 0.88nA. In the classic WTA, the output currents would show no response to the input current steps.

winner selection. The difference between the classic and adaptive WTA is that $\boldsymbol{M_4}$ and $\boldsymbol{M_5}$ are pFET single-transistor synapses; these floating-gate transistors null their inputs slowly over time. This enhancement results in the ability of each transistor to adapt to its input bias current. The adaptation is a result of the electron tunneling and hot-electron injection modifying the charge on the floating gate. The circuit is devised such that these are negative-feedback mechanisms; consequently, the output voltage always returns to the same steady-state voltage, determined by its bias current regardless of the DC input level.

Figure 6.4 shows characteristic traces from the two-input circuit. The winning node corresponds to the lowest voltage, which is reflected in that node's corresponding high output current. Looking at Fig. 6.4(a) we see that, as an input step is applied, the output current jumps and then begins to adapt to a steady-state value. When the inputs are nearly equal, the steady-state outputs are nearly equal; when the inputs are different, the steady state output is greater for the cell with the lesser input. In general, the input current change that is the largest after reaching the previous equilibrium becomes the new equilibrium. This additional decrease in V_1 would lead to an amplified increase in the other voltage since the losing stage roughly looks like an autozeroing amplifier with the common node as the input terminal. The extent to which the inputs do not equal this largest input is manifested as a proportionally larger input voltage. The other voltage would return to equilibrium by slowly, linearly decreasing in voltage due to the tunneling current. This process will continue until V_1 equals V_2. Note that in general that the inputs with lower bias currents have a slight starting advantage over the inputs with higher bias currents.

Figure 6.4(b) illustrates the advantage of the adaptive WTA over the classic WTA. In the classic WTA, the output voltage and current would not change throughout the experiment, but the adaptive WTA responds to changes in the input. The second input step does not evoke a response because there was not enough time to adapt to steady state after the previous step; but the next step immediately causes it to win.

Bibliography

[1] M. Mahowald and R. Douglas, "A silicon neuron," *Nature*, vol. 354, no. 6345, 1991, pp. 515-518.

[2] B.A. Minch, P. Hasler, C. Diorio and C. Mead, "A silicon axon," in G. Tesauro, D.S. Touretzky, and T.K. Leen, *Advances in Neural Information Processing Systems 7*, MIT Press, Cambridge, MA, 1995, pp. 739-746.

[3] P. Hasler, "Implementing practical neural networks in silicon," *Proceedings of the Wescon Conference*, SS-1, 1988, pp. 1-7.

[4] P. Hasler, *An Implementation of a Continuous-Time Trainable Neural Network*, Master's Thesis, Arizona State University, 1991.

[5] P. Hasler and L. Akers, "Circuit implementation of a trainable neural network using the generalized Hebbian algorithm with supervised techniques," *Proceedings of the International Joint Conference on Neural Networks*, Baltimore, vol. I, 1992, pp. 1565-1568.

[6] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, USA, vol. 79, 1982, pp. 2554-2558.

[7] J.J. Hopfield and D.W. Tank, "Neural computations of decisions in optimization problems," *Biological Cybernetics*, vol. 52, 1985, pp. 141-152.

[8] D.E. Rumelhart and J.L. McClellend, *Parallel Distributed Processing, Exporations in the Microstructure of Cognition*, MIT Press, Cambride, MA, 1986.

[9] R.P. Lippman, "An introduction to computing with neural nets," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 4, no. 2, 1987.

[10] M. Mahowald and C. Mead, "The silcon retina," *Scientific American*, vol. 264, no. 5, 1991, pp. 76-82.

[11] T. Delbruck, "Silicon Retina with correlation-based velocity-tuned pixels," *IEEE Transactions on Neural Networks*, vol. 4, no. 3, 1993, pp. 529-541.

[12] K. Boahen, A. Andreou, "A contrast-sensitive retina with reciprocal synapses," in J.E. Moody, *Advances in Neural Information Processing Systems 4*, Morgan Kaufman Publishers, San Mateo, CA, 1991

[13] A.G. Andreou, "Low power analog VLSI systems for sensory information processing," in B. Sheu, E. Sanchez-Sinencio, and M. Ismail, *Microsystems technologies for multimedia applications*, IEEE Press, Los Alamitos, CA, 1995.

[14] M. Mahowald, *An Analog VLSI Stereoscopic Vision System*, Kluwer Academic Publishers, Boston, MA, 1994.

[15] R.F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, 1988, pp. 1119-1134.

[16] L. Watts, D.A. Kerns, and R.F. Lyon, "Improved implementation of the silicon cochlea," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 5, 1992, pp. 692-700.

[17] J. Lazzaro and C. Mead, "A silicon model of auditory localization," *Neural Computation*, vol. 1, 1989, pp. 47-57.

[18] C. Mead, X. Arreguit, and J. Lazzaro, "Analog VLSI models of binaural hearing," *IEEE Transactions on Neural Networks*, vol. 2, 1991, pp. 230-236.

[19] M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable artificial neural network with 10240 'floating gate' synapses," *Proceedings of the International Joint Conference on Neural Networks*, Washington, D.C., vol. II, 1989, pp. 191-196.

[20] M. Cohen and A.G. Andreou, "Current-mode subthreshold MOS implementation of the Herault-Jutten autoadaptive network," *IEEE Journal of Solid State Circuits*, vol. 27, no. 5, 1992, pp. 714-727.

[21] B. Furman, J. White, and A.A. Abidi, "CMOS analog IC implementing the backpropagation algorithm," in *Abstracts of the First Annual INNS Meeting*, vol. 1, 1988, p. 381.

[22] R.G. Benson and D.A. Kerns, "UV-activated conductances allow for multiple time scale learning," *IEEE Transactions on Neural Networks*, vol. 4, no. 3, 1993, pp. 434-440.

[23] T. Serrano-Gotarredona, B. Linares-Barranco, and J.L. Huertas, "A real time clustering CMOS engine," in G. Tesauro, D.S. Touretzky, and T.K. Leen, *Advances in Neural Information Processing Systems 7*, MIT Press, Cambridge, MA, 1995.

[24] F.J. Pineda, G. Cauwenberghs, and R.T. Edwards, "Bangs, clicks, snaps, thuds, and whacks: an architecture for acoustic transient processing," in M.C. Moser, M.I. Jordan and T. Petsche, *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, MA, 1997.

[25] G. Graf, L. Jackel, R. Howard, B. Howard, B. Stranghn, J. Denker, W. Hubbard, D. Tennant, and D. Schwartz, "VLSI implementation of a neural network memory with several hundreds of neurons," *AIP Conference Proceedings, Snowbird 151*, 1986, pp. 182.

[26] Y. Tsividis and S. Satyanarayana, "Analogue circuits for variable synapse electronic neural networks," *Electronics Letters*, vol. 24, no. 2, 1987, pp. 1313-1314.

[27] F. Kub, K.K. Moon, I.A. Mack, and F.M. Long, "Programmable analog vector-matrix multiplier," *Journal of Solid State Circuits*, vol. 25, no. 1, 1990, pp. 207-214.

[28] Z. Czarnul, "Novel MOS resistive circuit for synthesis of fully integrated continuous-time filters," *IEEE Transactions on Circuits and Systems*, vol. 33, no. 2, 1986, pp. 277-281.

[29] S.T. Dupuie and M. Ismail, "High frequency CMOS transconductors," in C. Toumazou, F.J. Lidgey, and D.G. Haigh, *Analogue IC Design: the Current-Mode Approach*, Peter Peregrinus, London, 1990.

[30] P. Hasler and L.A. Akers, "A continuous-time synapse employing a multilevel dynamic memory" *Proceedings of the International Joint Conference on Neural Networks*, Seattle, vol. I, 1991, pp. 563-568.

[31] G. Cauwenberghs, C. Neugebauer, and A. Yariv, "An adaptive CMOS matrix vector multiplier for large scale analog hardware neural network applications," *Proceedings of the International Joint Conference on Neural Networks*, Seattle, vol. I, 1991, pp. 507-512.

[32] A. F. Murray, "Pulse arithmatic in VLSI neural networks," *IEEE Micro*, vol. 9, no. 6, 1989, pp. 64-74.

[33] T. Delbruck, "An electronic photoreceptor sensitive to small changes in intensity," in D.S. Touretzky, *Advances in Neural Information Processing Systems 1*, Morgan Kaufman, San Mateo, CA, 1988, pp. 720-727.

[34] T. Delbruck and C. Mead, "Photoreceptor circuit with wide dynamic range," *Proceedings of the International Circuits and Systems Meeting*, London, England, 1994.

[35] P. Hasler, C. Diorio, B.A. Minch, and C. Mead, "Single transistor learning synapses," in *Advances in Neural Information Processing Systems 7*, MIT Press, Cambridge, MA, 1995, pp. 817-824. Also at *http://www.pcmp.caltech.edu/anaprose/paul.*

[36] P. Hasler, C. Diorio, B.A. Minch, and C. Mead, "Single transistor learning synapses with long term storage," *Proceedings of the International Sympo-*

sium on Circuits and Systems, Seattle, vol. 3, 1995, pp. 1660-1663. Also at *http://www.pcmp.caltech.edu/anaprose/paul.*

[37] P. Hasler, B.A. Minch, C. Diorio, and C. Mead, "An autozeroing amplifier using *p*FET hot-electron injection," *Proceedings of the International Symposium on Circuits and Systems*, Atlanta, vol. 3, 1996, pp. 325-328. Also at *http://www.pcmp.caltech.edu/anaprose/paul.*

[38] B.A. Minch, C. Diorio, P. Hasler, and C. Mead, "Translinear circuits using subthreshold floating-gate MOS transistors," *Analog Integrated Circuits and Signal Processing*, vol. 9, no. 2, 1996, pp. 167-179. Also at *http://www.pcmp.caltech.edu/anaprose/bminch.*

[39] B.A. Minch, C. Diorio, P. Hasler, and C. Mead, "A νMOS soft-maximum current mirror," *Proceedings of the International Symposium on Circuits and Systems*, Seattle, vol. 3, 1995, pp. 2249-2252. Also at *http://www.pcmp.caltech.edu/anaprose/bminch.*

[40] B.A. Minch, C. Diorio, P. Hasler, and C. Mead, "The matching of small capacitors for analog VLSI," *Proceedings of the International Symposium on Circuits and Systems*, Atlanta, vol. 1, 1996, pp. 238-241. Also at *http://www.pcmp.caltech.edu/anaprose/bminch.*

[41] C. Diorio, P. Hasler, B.A. Minch, and C. Mead, "A complementary pair of four-terminal silicon synapses," *Analog Integrated Circuits and Signal Processing*, 1997.

[42] C. Diorio, P. Hasler, B.A. Minch, and C. Mead, "A single-transistor silicon synapse," *IEEE Transactions on Electron Devices*, vol. 43, no. 11, 1996, pp. 1972-1980.

[43] C. Diorio, S. Mahajan, P. Hasler, B.A. Minch, and C. Mead, "A high resolution non-volatile analog memory cell," *Proceedings of the International Conference of Circuits and Systems*, Seattle, vol. 3, 1995, pp. 2233-2236.

[44] A. Thomsen and M.A. Brooke, "A floating gate MOSFET with tunneling injector fabricated using a standard double-polysilicon CMOS process," *IEEE Electron Device Letters*, vol. 12, 1991, pp. 111-113.

[45] J. Mann and S. Gilbert, "An analog self organizing neural network chip," in D.S. Touretzky, *Advances in Neural Information Processing Systems 1*, Morgan Kaufmann, San Mateo, CA, 1988, pp. 739-747.

[46] B. Hochet, "Multivalued MOS memory for variable synapse neural networks," *Electronics Letters*, vol. 25, no. 10, 1989, pp. 669-670.

[47] B. Gilbert, "A precise four-quadrant multiplier with subnanosecond response," *IEEE Journal of Solid State Circuits*, vol. 3, no. 4, 1968, pp. 365-373.

[48] S. Sze, *Physics of Semiconductor Devices*, Wiley Interscience, New York, 1981.

[49] F. Masuoka, R. Shirota, and K. Sakui, "Reviews and prospects of non-volatile semiconductor memories," *IEICE transactions*, vol. E 74, no. 4, 1991, pp. 868-874.

[50] H.V. Tran, T. Blyth, D. Sowards, L. Engh, B.S. Nataraj, T. Dunne, H. Wong, V. Serin, T. Lam, H. Hazarian,and G. Hu, "A 2.5V 256-level non-volatile analog storage device using EEPROM technology," *Proceedings of IEEE International Solid-State Circuits Conference*, 1996, pp. 270-271.

[51] W. Shockley, "Problems related to p-n junctions in silicon," *Solid State Electronics*, vol. 2, no. 1, 1961, pp. 35-67.

[52] T.H. Ning, "Hot-electron emission from silicon in to silicon dioxide, " *Solid State Electronics*, vol. 21, 1978, pp. 273-282.

[53] S. Tam, P.K. Ko, and C. Hu, "Lucky-electron model of channel hot-electron injection in MOSFETs," *IEEE Transactions on Electron Devices*, 1984, pp. 1110-1125.

[54] J.J. Sanchez and T.A. DeMassa, "Review of carrier injection in the silicon / silicon-dioxide system," *IEE Proceedings-G,* vol. 138, no. 3, 1991.

[55] M.V. Fischetti and S.E. Lauz, "Monte carlo study of sub-band-gap impact ionization in small silicon field-effect transistor," *International Electron Device Meeting*, 1995, pp. 305-308.

[56] J.Y. Tang, and K. Hess, "Impact ionization of electrons in silicon," *Journal of Applied Physics*, vol. 54, no. 9, 1983, pp. 5139-5151.

[57] G.A. Baraff, "Distribution functions and ionization rates for hot-electrons in semiconductors," *Physical Review*, vol. 128, no. 6, 1962, pp. 2507-2517.

[58] G.A. Baraff, "Maximum anisotropy approximation for calculating electron distributions; application to high field transport in semiconductors," *Physical Review*, vol. 133, no. 1A, 1964, pp. A26-A33.

[59] D.K. Ferry, and C. Jacoboni, *Quantum Transport in Semiconductors*, Plenum Press, New York, 1992.

[60] E.M. Conwell, *High Field Transport in Semiconductors*, Academic Press, New York, 1967.

[61] M.V. Fischetti, S.E. Laux, and E. Crabbe, "Understanding hot-electron transport in silicon devices: Is there a shortcut?," *Journal of Applied Physics*, vol. 78, no. 2, 1995, pp. 1058-1087.

[62] P.A. Wolfe, "Theory of electron multiplication in silicon and germanium," *Physical Review*, vol. 95, no. 6, 1954, pp.1415-1420.

[63] L.V. Keldish, "Concerning the theory of impact ionization in semiconductors," *Soviet Physics JETP*, vol. 21, no. 6, 1967, pp. 1135-1145.

[64] K. Hess, "Phenomenological physics of hot carriers in semiconductors, " in D.K. Ferry, J.R. Barker, and C. Jacoboni, *Physics of Nonlinear Transport in Semiconductors*, Plenum Press, 1980, pp. 1-42.

[65] J. P. LeBurton and K. Hess, "Energy-diffusion equation for an electron gas interacting with polar optical phonons," *Phys. Review B*, vol. 26, no. 10, 1982, pp. 5623-5633.

[66] N. Sano and A. Yoshii, "Impact ionization rate near thresholds in Si," *Journal of Applied Physics*, vol. 75, 1994, pp. 5102-5105.

[67] Y.Kamakara, H. Mizuno, M. Yamaji, M. Morifuji, K. Taniguchi, C. Hamaguchi, T. Kunikiyo, and M. Takenaka, "Impact ionization model for full band Monte Carlo simulation," *Journal of Applied Physics*, vol. 75, 1994, pp. 3500-3506.

[68] J. Kolnik, Y. Wang, I.H. Oguzman, and K.F. Brennan, "Theoretical investigation of wave-vector-dependent analytical and numerical formulations of the interband impact-ionization transition rate for electrons in bulk silicon and GaAs," *Journal of Applied Physics*, vol. 76, 1994, pp. 3542-3551.

[69] J. Kevorkian and J. Cole, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, New York, 1981.

[70] G. B. Whittem, *Linear and Nonlinear Wave Propagation*, Wiley-Interscience, New York, 1973.

[71] H. Budd, "Path variable formulation of the hot carrier problem," *Physical Review*, vol. 158, no. 3, 1967, pp. 798-804.

[72] M.S. Shur and L.F. Eastman, "Balistic transport in semiconductor at low temperatures for low power, high speed logic," *IEEE Transations on Electron Devices*, vol. 6, no. 11, 1979, pp. 1677-1683.

[73] K. Hess, "Balistic electron transport in semiconductors," *IEEE Transactions on Electron Devices*, vol. 28, no. 8, 1981, pp. 937-940.

[74] Y. Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, New York, 1987.

[75] C. Chang, C. Hu, and R.W. Broaderson, "Quantum yield of electron impact ionization in silicon," *Journal of Applied Physics*, vol. 57, no. 2, 1985, pp. 302-309.

[76] C.H. Lee, U. Ravaioli, K. Hess, C. Mead, and P. Hasler, "Simulation of a long term memory device with a full bandstructure Monte Carlo approach," *IEEE Electron Device Letters*, vol. 16, no. 8, 1995, pp. 360-362.

[77] W. Quade, E. Scholl, and M. Rudan, "Impact ionization within the hydrodynamics approach to semiconductor transport," *Solid State Electronics*, vol. 36, no. 10, 1993, pp. 1493-1505.

[78] M. Lenzlinger and E.H. Snow, "Fowler–Nordheim tunneling into thermally grown $Si0_2$," *Journal of Applied Physics*, vol. 40, no. 1, 1969, pp.278–283.

[79] C. Mead, "Scaling of MOS technology to submicrometer feature sizes," *Journal of VLSI Signal Processing,* vol. 8, 1994, pp. 9-25.

[80] Y. Xu, *Electron Transport through Thin Film Amorphous Silicon — A Tunneling Study*, Ph.D. dissertation, Stanford University, 1992.

[81] P. Gray and R. Meyer, *Analysis and Design of Analog Integrated Circuits*, Wiley Interscience, New York, 1984.

[82] C. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, MA, 1989.

[83] P.S. Churchland and T.J. Sejnowski, *The Computational Brain*, MIT Press, Cambridge, MA, 1992.

[84] T.A. Fjeldly and M. Shur, "Threshold voltage modeling and the subthreshold regime of operation of short-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 40, no. 1, 1993, pp. 137-145.

[85] E.A. Vittoz, "Dynamic analog techniques," in Y. Tsividis and P. Antognetti, *Design of MOS VLSI Circuits for Telecommunications*, Prentice-Hall, Englewood Cliffs, NJ, 1985, pp. 145-170.

[86] Y. Leblebici and S. M. Kang, *Hot Carrier Reliability of MOS VLSI Circuits*, Kluwer Academic, Boston, 1993.

[87] Y. Tsividis, M. Banu, and J. Khaury, "Continuous-time MOSFET-C filters in VLSI," *IEEE Transactions on Circuits and Systems*, vol.33, no.2, 1986.

[88] R. Sarpeshkar, T. Delbruck, and C. Mead, "White noise in MOS transistors and resistors," *IEEE Circuits and Devices*, 1993, pp. 23-29. Also at *http://www.pcmp.caltech.edu/anaprose/rahul.*

[89] R. Sarpeshkar, R.F. Lyon, and C. Mead, "A low-power wide-linear-range transconductance amplifier," *Analog Integrated Circuits and Signal Processing*, 1997.

[90] E.H. Nicollian and J.R. Brews, *MOS Physics and Technology*, Wiley Interscience, New York, 1982.

[91] C. Hu, S. Tam, F. Hsu, P. Ko, T. Chan, and K. Terrill, "Hot-Electron-Induced MOSFET Degradation—Model, Monitor, and Improvement," *IEEE Transactions on Electron Devices*, vol. ED-32, no. 2, 1985, pp. 375-385.

[92] J. Lazzaro, S. Ryckebusch, M.A. Mahowald, and C. Mead, "Winner-take-all networks of O(n) complexity,"in D.S. Touretzky, *Advances in Neural Information Processing Systems 1*, Morgan Kaufmann, San Mateo, CA, 1988, pp. 703-711.

[93] W.F. Kruger, P. Hasler, B.A. Minch, and C. Koch, "An adaptive WTA using floating-gate technology," in M.C. Moser and M.I. Jordan and T. Petsche, *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, MA, 1997. Also at *http://www.pcmp.caltech.edu/anaprose/paul.*

Appendix A Appendix to Chapter 2

A.1 Derivation of Electron Characteristics

In this section, we solve the momentum angle characteristic. The characteristic equation for energy as a function of position ($E_1(z)$) is

$$\begin{aligned} \frac{d\zeta_1}{dz} &= q\mathcal{E}(z)\frac{1-\zeta^2}{\zeta E(z)}\frac{d\zeta_1}{dz} \\ &= q\mathcal{E}(z)\frac{1-(\zeta_1+\delta\zeta)^2}{(\zeta_1+\delta\zeta)(E_1(z)+\delta E)}. \end{aligned} \tag{A.1}$$

By using the chain rule for deriviates

$$\begin{aligned} \frac{d\zeta_1}{dE_1} &= \frac{d\zeta_1(z)}{dz}\frac{dz}{dE_1}, \\ &= \frac{q\mathcal{E}(z)}{q\mathcal{E}-\frac{E_R}{\zeta\lambda}}\frac{1-(\zeta_1+\delta\zeta)^2}{(\zeta_1+\delta\zeta)(E_1(z)+\delta E)}, \\ &= \frac{1-(\zeta_1+\delta\zeta)^2}{(E_1(z)+\delta E)(\zeta_1+\delta\zeta-\frac{E_R}{q\mathcal{E}(z)\lambda\zeta})}. \end{aligned} \tag{A.2}$$

Solving this equation assuming that $\mathcal{E}(z)$ is nearly constant in the region of interest results in the expression

$$\left(1-(\zeta_1+\delta\zeta)^2\right)\left(\frac{E_1(z)+\delta E}{E_a}\right)\left(\frac{1-(\zeta_1+\delta\zeta)}{1+(\zeta_1+\delta\zeta)}\right)^{\frac{E_R}{q\mathcal{E}\lambda}}=1, \tag{A.3}$$

where E_a is the parameter dependent upon the initial conditions.

A.2 Distribution Function with Impact Ionization Losses

A.2.1 Derivation of the $a(z, E)$ Equation

In this subsection, we want to simplify the Boltzman Transport Equation for $\zeta \approx 1$, (2.18), which is

$$\frac{\partial f}{\partial z} + \left(q\mathcal{E} - \frac{E_R}{\lambda}\right)\frac{\partial f}{\partial E} = F(T)\frac{E_R^2}{2\lambda}\frac{\partial^2 f}{\partial E^2} - \frac{f}{L(E)}.$$

When L(E) is not zero, we would like to factor out the solution of the lossless diffusion equation. To do this, we define f(z,E) in terms of two functions.

$$f(z, E) = g(z, E)a(z, E). \tag{A.4}$$

To substitute the functions g and a, we need the derivatives for $f(z, E)$:

$$\frac{\partial f}{\partial z} = a(z, E)\frac{\partial g}{\partial z} + g(z, E)\frac{\partial a}{\partial z}, \tag{A.5}$$

$$\frac{\partial f}{\partial E} = a(z, E)\frac{\partial g}{\partial E} + g(z, E)\frac{\partial a}{\partial E}, \tag{A.6}$$

$$\frac{\partial^2 f}{\partial E^2} = a(z, E)\frac{\partial^2 g}{\partial E^2} + 2\frac{\partial g}{\partial E}\frac{\partial a}{\partial E} + g(z, E)\frac{\partial^2 a}{\partial E^2}. \tag{A.7}$$

We define $g(z, E)$ as the solution to the original lossless diffusion equation

$$\frac{\partial g}{\partial z} + \left(q\mathcal{E} - \frac{E_R}{\lambda}\right)\frac{\partial g}{\partial E} = \frac{E_R^2}{2\lambda}F(T)\frac{\partial^2 g}{\partial E^2}. \tag{A.8}$$

The solutions for the previous subsection continue to apply above the impact ionization threshold, and are simply attenuated by the electron impact ionization by $a(z, E)$. With this definition of $g(z, E)$, the attenutation term, $a(z, E)$, solves

$$\frac{\partial a}{\partial z} + \left(q\mathcal{E} - \frac{E_R}{\lambda} - F(T)\frac{E_R^2}{\lambda}\frac{\frac{\partial g}{\partial E}}{g}\right)\frac{\partial a}{\partial E} = F(T)\frac{E_R^2}{2\lambda}\frac{\partial^2 a}{\partial E^2} - L(E)a, \tag{A.9}$$

which is roughly a Boltzman Transport equation for a constant boundary condition for all z at $E = E_{th}$.

A.2.2 Simplification of $a(z, E)$ under Electric Fields

In this subsection, we justify the simplifications of (2.31) which result in (2.37). By scaling the energy after E_{th} by $q\mathcal{E}(d)l_{ion}$ which we define as E_s we transform the following terms as

$$F(T)\frac{E_R^2}{\lambda}\frac{\frac{\partial g}{\partial E}}{g}\frac{\partial a}{\partial E} \rightarrow F(T)\frac{E - E_1(z)}{qV(z)}\frac{\partial a}{\partial x}, \tag{A.10}$$

$$F(T)\frac{E_R^2}{2\lambda}\frac{\partial^2 a}{\partial E^2} \rightarrow \left(\frac{E_R}{q\mathcal{E}(d)l_{ion}}\right)^2 F(T)\frac{l_{ion}}{2\lambda}\frac{\partial^2 a}{\partial x^2}, \tag{A.11}$$

$$\left(q\mathcal{E} - \frac{E_R}{\lambda}\right)\frac{\partial a}{\partial E} \rightarrow \frac{q\mathcal{E} - \frac{E_R}{\lambda}}{q\mathcal{E}(d)l_{ion}}\frac{\partial a}{\partial x}. \tag{A.12}$$

The diffusion term is a second order perturbation, and the $\frac{\frac{\partial g}{\partial E}}{g}$ term is a small quantity in the ranges of interest. The small diffusion term means that a is not a rapidly moving function in energy compared E_R. Since $q\mathcal{E}\lambda$ will be much larger than E_R in the hot-electron injection or impact ionization regions, the diffusion term has a second order effect. The derivatives in $a(z, E)$ are not large enough to make the second order terms have a first order effect in the original equation.

If we can ignore these terms, we reduce the problem to

$$\frac{\partial a}{\partial z} + \left(q\mathcal{E} - \frac{E_R}{\lambda}\right)\frac{\partial a}{\partial E} = -L(E)a. \tag{A.13}$$

By substituting this solution back into (A.9) we get less than a 10 percent change in $L(E)$. This error would be smaller for lower substrate dopings. If explicitly solving this equation, one gets parrallel characteristics or flow lines where $a(z, E)$ changes very slowly between each of the characteristics, which is expected since we have small percentage changes in $\mathcal{E}(z)$ near the drain edge. This implies that the derivative in z is small.