Neurally Inspired Silicon Learning: From Synapse Transistors to Learning Arrays

Thesis by

Chris Diorio

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

California Institute of Technology
Pasadena, California
1997
(Defended 13 May 1997)

Acknowledgments

How strange is the lot of us mortals! Each of us is here for a brief sojourn; for what purpose he knows not, though he sometimes thinks he senses it. But without deeper reflection one knows from daily life that one exists for other people—first of all for those upon whose smiles and well-being our own happiness is wholly dependent, and then for the many, unknown to us, to whose destinies we are bound by the ties of sympathy. A hundred times every day I remind myself that my inner and outer life are based on the labors of other men, living and dead, and that I must exert myself on order to give in the same measure as I have received and am still receiving.

—Einstein, 1931

To Christina, and her simple question: "Daddy, will you be home tonight?" I hope now to be able to answer always yes.

To Alexia: Your inquisitiveness is my joy and my inspiration—ask of everything "*why?*"

To Mary-Ellen: Your unwavering confidence and love is, and always will be, my strength and my life.

To Carver: You opened the door to a new world for me—I am forever grateful.

To Mary and Paul: I can only admire your courage.

To Mom and Dad and Philip and Elisa: Thank you for your encouragement.

To Brad and Paul: The years of collaboration were invaluable—thank you.

To Sunit and Sanjoy: You are my inspiration to always do better.

To Buster and Shih-Chii and Tobi and Rahul and Lena and Jeff: Thank you for the shared times—the good, the bad, the help, the criticism, and the mutual support.

To Yaser and Gilles and Christof and Demetri and Rod and Joel: Thank you for welcoming me into the future.

To Helen and Candace and Donna and Jim and Calvin: I wish you always the best.

To Lyn: You yet may get me to write good...

My thanks for support from: TRW Inc.; the Office of Naval Research; the Advanced Research Projects Agency; the Beckman Hearing Institute; the Center for Neuromorphic Systems Engineering as a part of the National Science Foundation's Engineering Research Center Program; and the California Trade and Commerce Agency, Office of Strategic Technology.

And finally, my thanks to the MOSIS organization, for offering a chip fabrication service that makes student research possible.

Abstract

A computation is an operation that can be performed by a physical machine. We are familiar with digital computers: Machines based on a simple logic function (the binary NOR) and optimized for manipulating numeric variables with high precision. Other computing machines exist: The neurocomputer, the analog computer, the quantum computer, and the DNA computer all are known. Neurocomputers—defined colloquially as computing machines comprising nervous tissue—exist; that they are computers also is certain. Nervous tissue solves ill-posed problems in real time. The principles underlying neural computation, however, remain for now a mystery.

I believe that there are fundamental principles of computation that we can learn by studying neurobiology. If we can understand how biological information-processing systems operate, then we can learn how to build circuits and systems that deal naturally with real-world data. My goal is to investigate the organizational and adaptive principles on which neural systems operate, and to build silicon integrated circuits that compute using these principles. I call my approach *silicon neuroscience:* the development of neurally inspired silicon-learning systems.

I have developed, in a standard CMOS process, a family of single-transistor devices that I call *synapse transistors.* Like neural synapses, synapse transistors provide nonvolatile analog memory, compute the product of this stored memory and the applied input, allow bidirectional memory updates, and simultaneously perform an analog computation and determine locally their own memory updates. I have fabricated a synaptic array that affords a high synapse-transistor density, mimics the low power consumption of nervous tissue, and performs both fast, parallel computation and slow, local adaptation. Like nervous tissue, my array simultaneously and in parallel performs an analog computation and updates the nonvolatile analog memory.

Although I do not believe that a single transistor can model the complex behavior of a neural synapse completely, my synapse transistors do implement a local learning function. I consider their development to be a first step toward achieving my goal of a silicon learning system.

Contents

Figures

Tables

Preface: From Neurobiology to Silicon

My goal is to build electronic systems that employ the computational and organizational principles used in the nervous systems of living organisms. Nervous systems solve, in real time, ill-posed problems in image and speech processing, motor control, and learning; they do so in ways that we, as scientists and engineers, do not understand. There are fundamental principles that we can learn from neurobiology about a different and—on poorly conditioned data—vastly more efficient form of computation.

I believe that there is nothing that is done in the nervous system that we cannot emulate with electronics, once we understand the principles of neural information processing. Although nervous tissue solves problems that we do not know how to solve, it does so using an underlying device physics that we know and understand. A similar device physics underlies the semiconductor electronics that we employ to build our digital computers.

In both integrated circuits and nervous tissue, information is manipulated principally on the basis of charge conservation. In semiconductor electronics, electrons are in thermal equilibrium with their surroundings; their energies are Boltzmann distributed. In nerve tissue, ions are in thermal equilibrium with their surroundings; their energies also are Boltzmann distributed. In semiconductor electronics, we erect energy barriers to contain the electronic charge, by using the work-function difference between silicon and silicon dioxide, or the energy barrier in a *pn* junction. The nervous system erects similar energy barriers to contain its electronic charge, by using lipid membranes in an aqueous solution. In both systems, when the height of the energy barrier is modulated, the resulting current flow is an exponential function of the applied voltage. Both systems use this principle to produce devices that exhibit signal gain. Transistors use populations of electrons to change their channel conductance, in much the same way that neurons use populations of ionic channels to change their membrane conductance.

I believe that the disparity between the computations that can be done by a digital computer and those that can be done by the nervous system is a consequence of the way that the underlying physics is used to effect the computation. The state variables in both electronic and nervous sys-

tems are analog. They are represented in electronic systems by electric charge, and in nervous systems by electric charge or by chemical concentrations. The mechanisms by which each system manipulates its state variables to do computation, however, are vastly different. In a digital computer, we ignore most of the available states in favor of the two binary-valued endpoints: We achieve noise immunity at the expense of dynamic range. The nervous system retains the analog dynamic range, achieving noise immunity by adjusting the signal-detection threshold adaptively. Digital machines quantize their analog inputs, and use restoring logic at every computational step. Nervous systems perform primarily analog computations, and quantize the computed result.

Unfortunately, we do not know what computational primitives neural systems use, how they represent information, or what their organizing principles are. However, because semiconductor electronics allows us to apply, at a high level of integration, a device physics similar to that used by neural tissue, I conclude that we should be able to build electronic circuits that mimic the computational primitives of nervous systems, and that we should be able to use these circuits to explore the organizational principles employed by neurobiology. I call my approach *silicon neuroscience:* the development of neurally inspired silicon learning systems.

My predecessors began these investigations by modeling two of the sensory organs available to neural systems: the retina and the cochlea [1, 2, 3]. The silicon retina and cochlea are now well developed, and mimic a portion of the sensory preprocessing performed by living organisms. Other researchers have made substantial progress in modeling the motor-control systems employed by living organisms [4, 5]. My colleagues Paul Hasler, Bradley Minch, and I are now beginning to model what is perhaps the most remarkable aspect of living organisms: their abilities to adapt and to learn [6–9].

The nervous system has mechanisms for long-term memory and for learning, including synaptic plasticity and neuronal growth [10]. Semiconductor electronics also has mechanisms for long-term memory—in particular, the nonvolatile EEPROM devices. I have adapted the floating-gate technology used in digital EEPROM devices to allow nonvolatile analog storage and to perform a local learning function, and I have done so using a standard CMOS process. I have developed a family of single-transistor devices that I call *synapse transistors* [11–14]; these devices, like neural synapses, implement long-term nonvolatile analog memory, allow bidirectional memory updates, and can learn from an input signal without interrupting the ongoing computation. My synapse transistors also compute the product of their stored analog memory and the applied input. Although I do not believe that a single device can model the complex behavior of a neural synapse completely, my synapse transistors do implement a local learning function.

Neurons are the nervous system's primary computing elements. A typical neuron is markedly unlike a typical logic gate: It possesses on average 10,000 synaptic inputs, and a similar number of outputs. Its stored memory is contained in the pattern and strength of the analog synapses that connect it to other neurons. Nervous systems use vast numbers of synapses to effect their computations: In neocortical tissue, the synapse density is roughly 3×10^8 synapses per cubic millimeter [15].

I will use vast numbers of silicon synapses to model nervous tissue: Synapse-transistor arrays, fabricated in a standard CMOS process, afford a high device density, mimic the low power consumption of neural synapses, and perform both parallel computation and local adaptation. Like neural tissue, arrays of silicon synapse transistors simultaneously perform an analog computation and update their nonvolatile analog weight values.

I believe that, if we can understand the principles on which biological information-processing systems operate, we can build circuits and systems that deal naturally with real-world data. My goal, therefore, is to consider the computational principles on which neural systems operate, and to model and understand those principles in the silicon medium. I consider the development of synapse transistors to be a first step toward achieving my goal.

1 M. A. Mahowald and C. Mead, "The silicon retina," *Scientific American,* vol. 264, no. 5, pp. 76–82, 1991.

2 K. Boahen, "A retinomorphic vision system," *IEEE Micro,* vol. 16, no. 5, pp. 30–39, 1996.

3 L. Watts, D. A. Kerns, R. F. Lyon, and C. A. Mead, "Improved implementation of the silicon cochlea," *IEEE J. Solid-State Circuits*, vol. 27, no. 5, pp. 692–700, 1992.

4 S. P. DeWeerth, L. Nielsen, C. A. Mead, and K. J. Astrom, "A simple neuron servo," *IEEE Tran. Neural Networks*, vol. 2, no. 2, pp. 248–251, 1991.

5 T. Horiuchi, T. Morris, C. Koch, and S. P. DeWeerth, "Analog VLSI circuits for attention-based visual tracking," *Advances in Neural Information Processing Systems 9,* MIT Press (in press).

6 P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single transistor learning synapses," in *Advances in Neural Information Processing Systems 7*, G. Tesauro, D. Touretzky, and T. Leen, eds., MIT Press: Cambridge, MA, pp. 817–824, 1995.

7 P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single transistor learning synapses with long term storage," in *Proc. IEEE Intl. Symp. on Circuits and Systems*, Seattle, WA, vol. 3, pp. 1660–1663, 1995.

8 C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A single-transistor silicon synapse," *IEEE Trans. Electron Devices,* vol. 43, no. 11, pp. 1972–1980, 1996.

9 C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A complementary pair of four-terminal silicon synapses," *Analog Integrated* Circuits *and Signal Processing* (in press).

10 C. Koch, "Computation and the single neuron," *Nature,* vol. 385, no. 6613, pp. 207–210, Jan. 16, 1997.

11 C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A semiconductor structure for long term learning," U.S. Patent No. 5,627,392, issued 6 May, 1997.

12 C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A three-terminal silicon synaptic device," provisional patent application submitted to the U.S. Patent Office on 15 November, 1995, and assigned serial no. 60/006,795; utility patent application submitted to the U.S. Patent Office on 25 July, 1996.

13 C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "Hole impact-ionization method of hot-electron injection and a four-terminal *p*FET semiconductor structure for long-term learning," provisional patent application submitted to the U.S. Patent Office on 29 April, 1996, and assigned serial no. 60/016,464; utility patent application submitted to the U.S. Patent Office on 22 April, 1997.

14 C. Diorio and C. Mead, "A *p*MOS analog EEPROM cell," provisional patent application submitted to the U.S. Patent Office on 24 July, 1996, and assigned serial no. 60/022,360.

15 R. Douglas, "Rules of thumb for neuronal circuits in the neocortex," *Notes for the Neuromorphic aVLSI Workshop,* Telluride, CO, 1994.

Chapter 1

Neurobiology, Neural Networks, and Synapse Transistors

1.1 Neurobiology

The term *synapse* celebrates its centennial birthday this year: Sherrington first introduced the term in 1897 to designate a functional junction between nerve cells [1]. Over the past 100 years, there has emerged an increasing appreciation of the subtlety of form and function of these minute junctions: Synapses now are believed to be central to nervous function [2]. In the past decade, an avalanche of new information about synaptic junctions has been revealed by myriad electronic, chemical, and pharmacological tools; this information leads us to believe that significant information processing arises at or near synaptic interfaces.

A synapse is an anatomically distinct junction between two cells, at least one of which is a neuron [3]. The morphology of neuronal synapses is reasonably well known [4]. The presynaptic and postsynaptic membranes are separated by a synaptic cleft roughly 200Å wide. Communication between the cell membranes typically is chemically mediated, is unidirectional, can be excitatory or inhibitory, and can be locally sustained or reversed by repetitive stimuli. Action potentials (binary-valued voltage spikes) arriving at a presynaptic terminal cause, in most cases, the release of a chemical neurotransmitter; this neurotransmitter diffuses across the synaptic cleft to cause a potential change in the postsynaptic membrane. This postsynaptic potential is graded, differing from the binary-valued action-potential input. Changes in the postsynaptic membrane's potential and conductance are the variables that underlie synaptic communication.

The synaptic density and connectivity of mammalian nervous tissue is astounding. Neocortical tissue comprises roughly 3×10^8 synapses per mm^3 [5]. A typical neocortical pyramidal neuron

may communicate with 10,000 other neurons, usually at a single synaptic junction [5]. Neuronal signaling is parallel and asynchronous; consequently, postsynaptic potentials are responsible for the communication and transformation of immense quantities of information.

Synaptic facilitation is defined as an increase in a synapse's postsynaptic response as a result of prior stimulation. Usage-dependent synaptic facilitation is well known in nervous tissue; it arises usually as a result of presynaptic stimulation that is coincident with postsynaptic depolarization [6, 7, 8]. Post-tetanic potentiation (PTP), defined as a short-term increase in postsynaptic response, typically lasts for minutes. Long-term potentiation (LTP), defined as a long-term increase in postsynaptic response, can last for days or for weeks. Finally, long-term depression (LTD), defined as a long-term decrease in synaptic efficacy, has recently been observed in nervous tissue [9]. These effects are nonlinear with presynaptic activity, and the temporal order of the presynaptic and postsynaptic stimuli has been shown to be critical. If the presynaptic input precedes the postsynaptic depolarization, then the synapse undergoes LTP; if the timing is reversed, then the synapse undergoes LTD [10].

The diversity of synaptic and neuronal morphologies and the narrow boundaries of our present knowledge warn against excessive generalization about nervous tissue. The range of synaptic properties includes, at a minimum, polarized and nonpolarized transmission (both chemical and electrical), excitation, inhibition, graded potentials, variable membrane conductances, delay, summation, facilitation, and thresholding; most of these parameters are nonlinear and time variable [3, 4]. This mixture is discouragingly complicated—but this neurophysiology is the one with which we live, and this computational medium is the one that we hope to understand.

1.2 The Neural-Network Revolutions

The study of neurally inspired networks rests on the insights of the neuroanatomist Santiago Ramón y Cajal, who in 1911 inferred that the brain could store information and make associations by modifying connections between nerve cells [11], and the psychologist Donald Hebb, who in 1949 suggested that such modifications should take place if, and only if, the connected cells were active simultaneously [12]. The principle that networks can form internal representations by means of associations—by forming, strengthening, or pruning synaptic connections on the basis of mutual information—remains common to nearly all present-day neural-network models.

The neural-networks field began in the 1940s, and has been marked by three cycles of enthusiasm and subsequent skepticism. The first, a result of McCulloch and Pitts' discovery of emergent behavior from a distributed arrangement of simple neuronlike elements [13], was followed

by a second in the 1960s with Rosenblatt's perceptron [14]. The third occurred in the 1980s, when Hopfield used energy surfaces to describe network convergence [15], and Rumelhart and associates developed the backpropagation method for training multilayer networks [16]. As the term *neural network* implies, a primary goal is modeling the computational attributes of real neuronal networks. Unfortunately, it is precisely this goal that has led to the observed cycles: Each advance yields not the desired result—a computing machine modeled after the brain—but rather further insights into the enormity of the problem. Mead proposed that we apply our most advanced technology—silicon VLSI—as a tool to engineer neurally inspired systems; the term *neuromorphic engineering* derives from this proposal [17].

Neural-network researchers construct models—mathematical, software, or hardware—that exhibit computational similarities to nervous tissue. Throughout the history of neural-network modeling, the locus of interest has been on synaptic weights. McCulloch and Pitts employed neurons with fixed synaptic weights. Although their neurons were simple, they showed that networks of such neurons could perform many of the logic functions used in digital computers.

Rosenblatt's perceptron was the first viable neuron model with locally computed weight updates [14]; the perceptron learning rule owed its origins to the Hebb rule. Unfortunately, perceptron weights converge only for linearly separable problems, as noted by Minsky and Papert in their denouement of perceptrons in 1969 [18]. Although the problem of linear inseparability could be tackled by the addition of a hidden layer to the network, at that time there was no method for computing the weight values for the internal nodes.

The discovery of a backpropagation-of-errors learning rule for determining the synaptic weights of hidden units locally [16], using a variant of Hebb's rule, renewed interest in neural networks in the 1980s. Hopfield contributed to the excitement by introducing energy surfaces and free-energy minimization to the neural-network field; he showed how changing the synaptic weights changed a network's energy-surface profile, causing the network to converge to different solutions for a given input [15].

Throughout a half-century of neural-networks research, the dominant viewpoint within the neural-networks community remained that memory is stored in the synaptic weights. The first physiological evidence for synaptic modification in real neurons was published in 1973, when researchers modified synaptic strengths in the hippocampus [6]. There is now a great deal of evidence linking hippocampal LTP to memory formation; for this reason, the dominant viewpoint within the neuroscience community is that synaptic plasticity is the most plausible model for memory formation in the brain [19]. Because neural-network and neurophysiological research both indicate that synaptic plasticity offers the best model for the formation of associations,

memories, and complex representations in neural-computing networks, I have chosen to begin my research at the level of the synaptic junction.

1.3 Synapse Transistors

A neural network is a collection of interacting elements that operate locally and in parallel, and that as a whole exhibit emergent properties. Emergent systems represent a new paradigm in the engineering world: The systems are so complex that detailed engineering is neither possible nor desirable. We do not yet have the ability to fabricate computing networks in the medium used by neurobiology—hydrocarbons and aqueous solutions—but we can investigate such computing networks using the physics of silicon and metal. Following Mead's vision, I believe that neurobiology and silicon VLSI afford us an opportunity to learn how to build systems that design themselves.

There are essentially two complementary schools of thought in the development of neurally inspired hardware and systems. In the first, the intent is to reproduce physiological phenomena to increase our understanding of the nervous system [20]. In the second, the intent is to use a manageable subset of neural properties to investigate emergent behavior in networks of neuronlike elements [21]. My research falls into the second category: My goal is to build neurally inspired computing machines in silicon.

Researchers in this second category make the tacit assumption that reproducing many neurophysiological details is secondary to understanding the collective behavior of nervous tissue. Although this assumption may eventually prove erroneous, for now our impoverished tools and technology (relative to those of biology) make it necessary. Consequently, my work relates to neuroscience not at the level of neuroanatomical realism, but rather as a representation: I believe that if I can mimic, in silicon, a sufficient subset of the fundamental properties of nervous tissue, then I will be able to build neurally inspired computing machines that exhibit behavior analogous to that of nervous systems.

Silicon integrated-circuit processing allows us to make large numbers of nominally identical circuit elements on a single integrated chip. The synaptic density of nervous tissue, however, is enormous, even when compared with our most advanced silicon technology. To maximize the synaptic density in silicon-learning systems, I believe that I must incorporate the computational, memory-storage, and learning features of neural synapses within a single transistor. To ensure manageable system-level power consumption, I must make the power consumed by these transistors similar to that consumed by neural synapses. In addition, the output from these transistors should be a graded, analog signal, consistent with the postsynaptic response of neural synapses. I

cannot use digital logic to achieve these goals, but must instead use innate analog state variables available from the silicon-MOS physics. The nine properties that I have chosen to incorporate into my silicon-synapse model are:

1. Synapses possess nonvolatile analog memory.
2. When a synapse is not learning, the memory is nonvolatile.
3. When a synapse is learning, the memory updates can be bidirectional.
4. The synaptic output is the product of the input signal and the stored memory value.
5. Synaptic communication and memory updates can occur simultaneously.
6. Memory updates can vary with both the input signal and the stored memory value.
7. Synapses are small, and can be densely packed.
8. Synapses operates off a single-polarity supply.
9. Synaptic power consumption is small.

My hope is that systems fabricated from artificial silicon synapses will exhibit features and behavior similar to those exhibited by nervous tissue.

1.3.1 Existing Synapse Devices

Many researchers have built synapselike devices in silicon. The range of approaches is extensive, including complex, mixed analog–digital circuits [22]; devices or circuits that employ capacitive memory storage [23]; UV-writeable floating-gate devices [24]; electrically writeable floating-gate devices [25, 26]; and many others. The synapse-weight updates may be computed locally, using a variant of Hebb's rule [27], or they may be computed off-chip, and written serially to the local storage element [25].

The advantages of using floating-gate devices for memory storage are well known [28, 29]. The memory is nonvolatile, the charge storage is inherently analog, and the devices are compact. Other researchers have fabricated floating-gate silicon synapses [30, 31], both using electrically erasable, programmable, read-only memory (EEPROM) technology, and using conventional CMOS technology. EEPROM processing is highly specialized, and the devices are optimized for binary-valued (digital) storage [32]; consequently, EEPROM devices have not demonstrated the entirety of analog functions that I desire in a silicon synapse—in particular, they have not demonstrated locally computed weight updates. Floating-gate devices fabricated in conventional CMOS processes have likewise failed to demonstrate the analog functions that I desire in a silicon synapse, primarily because writing the memory requires high voltages and specialized device layouts. Consequently, there have been no single-transistor floating-gate circuit elements that combine analog storage, low power consumption, and locally computed weight updates.

1.3.2 Electrically Writeable Floating-Gate MOS Technology

The earliest experiments involving charge storage on the insulated gate of a field-effect transistor took place in the 1960s [33, 34, 35]. The goal at that time was to find a replacement for magnetic (core) memories. Digital EEPROM chips realized commercial success roughly a decade later. In the 1980s, Alspector and Allen suggested that the memory function in a neural network could be realized by using the analog charge on a floating gate to alter a transistor's threshold voltage [28]. Many researchers have subsequently investigated synapse devices employing electrically writeable floating-gate technology [36, 37].

The programming characteristics of floating-gate devices are quite variable, even for nominally identical transistors on the same chip [38]. For digital memories, these device variations can be compensated by applying excess charge in the write or erase processes. For analog memories, these device variations necessitate the use of feedback to ensure accurate memory writes. Various feedback mechanisms have been tried: Most employ either multi-step, iterative writes [38]; or single-step, open-loop writes with frequent calibration to compensate device mismatch and oxide degradation [25]. Neither approach supports locally computed, parallel weight updates in arrays of floating-gate devices.

I have developed four floating-gate MOS devices (two *p*FETs and two *n*FETs), in a standard CMOS process, that permit simultaneous memory reading and writing. In all four devices, the source and oxide currents are simultaneous, analog, and continuously valued. Consequently, I can use continuous analog feedback to write the charge on the floating gate carefully(see Section 3.2.5), and I can use continuous feedback to implement either a weight-update rule or a weight constraint in a silicon-learning system (see Section 4.2).

My devices comprise a single transistor, possess nonvolatile analog weight storage, permit simultaneous source and oxide currents, compute locally the product of their stored weight and the applied input, and decide locally their own weight updates. I store the weight as charge on the floating gate (see Figure 1.1). To mimic the power consumption of neural synapses, I operate the transistors in the subthreshold regime [39]. I modify the floating-gate charge bidirectionally by using Fowler–Nordheim (FN) tunneling [40] to remove electrons from the floating gate, and by using hot-electron injection [41] to add electrons to the floating gate.

In the FN-tunneling process (see Figure 1.2), a potential difference between the tunneling implant and the floating gate reduces the effective thickness of the gate-oxide barrier, facilitating electron tunneling from the floating gate, through the SiO_2 barrier, into the oxide conduction band. These electrons then are swept over to the tunneling implant by the tunneling-implant–to–

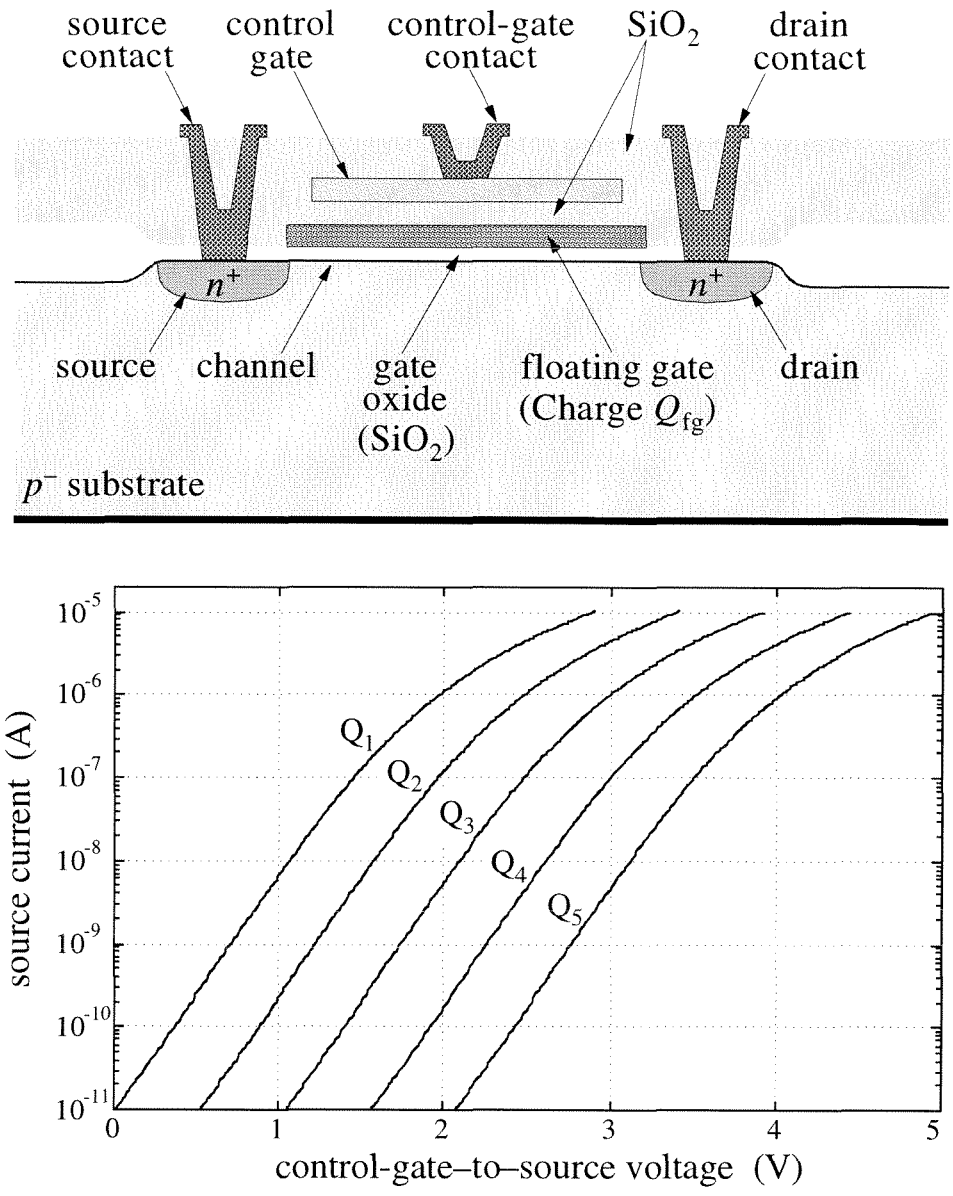


Figure 1.1 A floating-gate MOSFET stores a weight. I define the weight, W, in terms of the floating-gate charge, Q_{fg}, by $W=\exp(Q_{fg}/Q_T)$ (see Section 1.3.2.1). I employ both n-type and p-type floating-gate MOSFETs in my synapse transistors. I show an n-type MOSFET here; the p-type device is functionally similar. I apply signal inputs to the poly2 control gate, which couples capacitively to the poly1 floating gate. I can choose source current, drain current, or channel conductance as the output; because these quantities all are related [39], I use only source current in my subsequent analysis. I typically use subthreshold source currents, for reasons that I discuss in Section 1.3.2.1. From the control gate's perspective, changing Q_{fg} shifts the transistor's threshold voltage (bidirectionally).

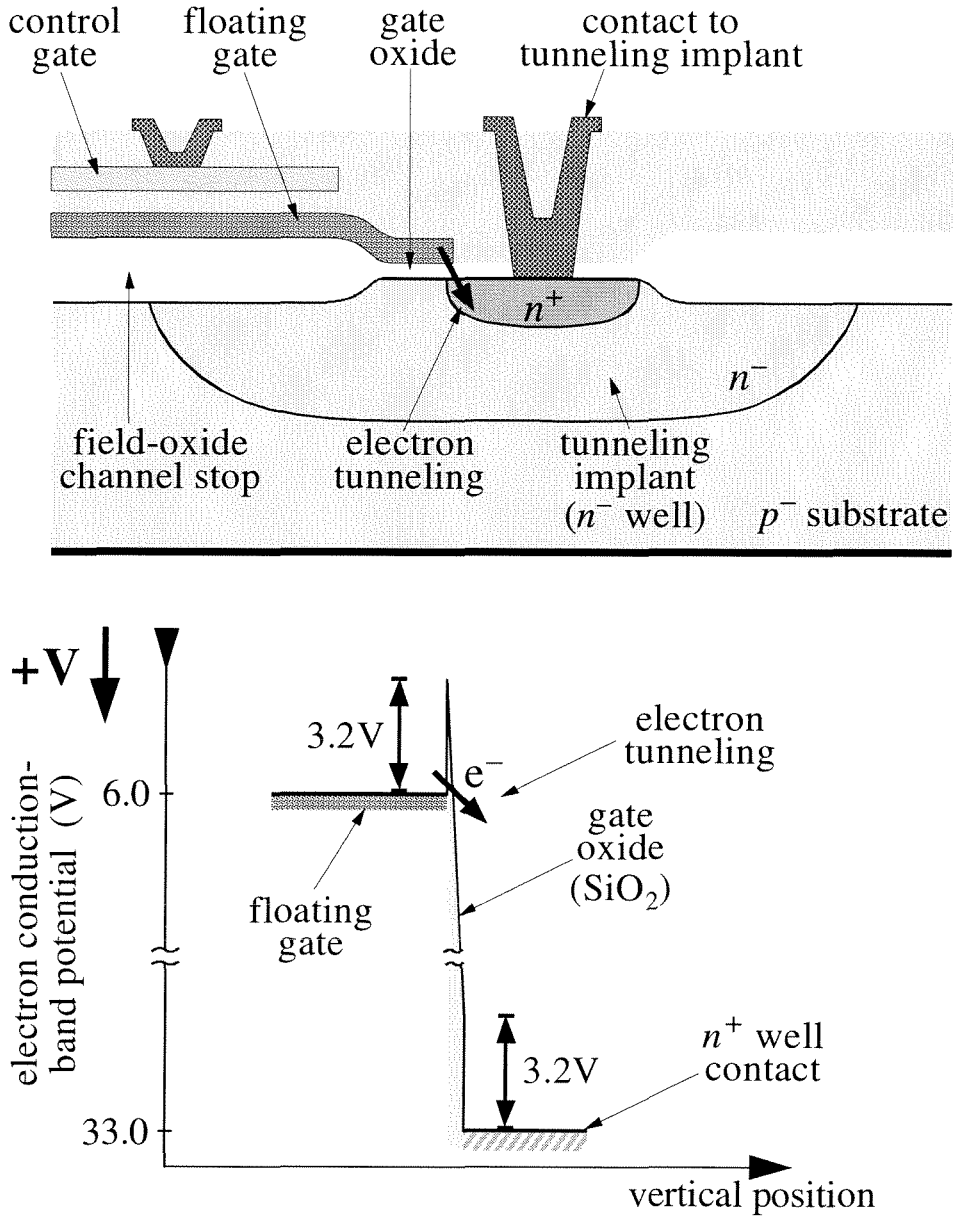


Figure 1.2 Electron tunneling increases the weight. I tunnel electrons from the floating gate, through the gate oxide, to the n^+ tunneling implant. All my devices require large tunneling voltages, because the gate-oxide thickness in the 2μm process that I use ranges from 350–450Å. Consequently, I must surround the n^+ tunneling implant with an n^- well, to prevent breakdown of the reverse-biased *pn* junction from the *n*-type tunneling implant to the *p*-type substrate. More modern processes have thinner oxides and therefore lower tunneling voltages; consequently, in future devices I anticipate replacing the n^- well with a graded junction [42].

floating-gate oxide electric field. In the hot-electron injection process (see Figure 1.3), electrons accelerate to high energies in the transistor's drain-to-channel electric field. A fraction of these electrons scatter upward into the gate oxide, and inject over the Si–SiO_2 work-function barrier into the oxide conduction band. These electrons then are swept over to the floating gate by the floating-gate–to–drain oxide electric field.

I fabricate all my devices in a standard 2μm *n*-well CMOS process (with NPN option) available from MOSIS (the 2μm Orbit process). Unlike digital EEPROMs, my devices require no special process-fabrication steps; I simply modify the implant and gate locations and geometries, within the standard process, to allow simultaneous channel and oxide currents. In addition, because I use a standard process, my devices can be integrated with conventional digital or analog MOS circuits elements.

1.3.2.1 Floating-Gate Transistors Store a Weight

I choose source current as the transistor output. I apply signal inputs to the poly2 control gate, which in turn couples capacitively to the poly1 floating gate. By operating the transistor in the subthreshold regime, I obtain three benefits. First, subthreshold channel currents ensure low power consumption—typically less than 100nW per device. Second, because the source current in a subthreshold MOSFET is an exponential function of the gate voltage, small changes to the floating-gate charge shift the transistor's threshold voltage measurably. Third, the output is the product of a stored weight and the applied input:

$$I_s = \mathrm{I}_0 e^{\frac{\kappa V_{fg}}{\mathrm{U_t}}} = \mathrm{I}_0 e^{\frac{\kappa\left(Q_{fg}+\mathrm{C_{in}}V_{in}\right)}{\mathrm{C_T U_t}}} = \mathrm{I}_0 e^{\frac{Q_{fg}}{\mathrm{Q_T}}} e^{\frac{\kappa' V_{in}}{\mathrm{U_t}}} \tag{1.1}$$

$$= W \mathrm{I}_0 e^{\frac{\kappa' V_{in}}{\mathrm{U_t}}} \tag{1.2}$$

where I_s is the source current, $\mathrm{I_0}$ is the pre-exponential current, κ is the coupling coefficient from the floating gate to the channel, Q_{fg} is the floating-gate charge, $\mathrm{C_T}$ is the total capacitance seen by the floating gate, $\mathrm{U_t}$ is the thermal voltage kT/q, $\mathrm{C_{in}}$ is the input (poly1 to poly2) coupling capacitance, V_{in} is the control-gate voltage, $\mathrm{Q_T}\equiv\mathrm{C_T U_t}/\kappa$, $\kappa'\equiv\kappa\mathrm{C_{in}}/\mathrm{C_T}$, $W\equiv\exp(Q_{fg}/\mathrm{Q_T})$, and, for simplicity, the source potential is assumed to be ground ($\mathrm{V_s}=0$). The weight W is the learned quantity: Its value derives from the floating-gate charge, which can change with synapse use.

1.3.2.2 Floating-Gate Transistors Multiply

The transistor output is the product of W and the source current of an idealized MOSFET that has a control-gate input V_{in} and a coupling coefficient κ′ from the control gate to the channel (see Eqn. (1.2)). Consequently, subthreshold floating-gate MOSFETs multiply.

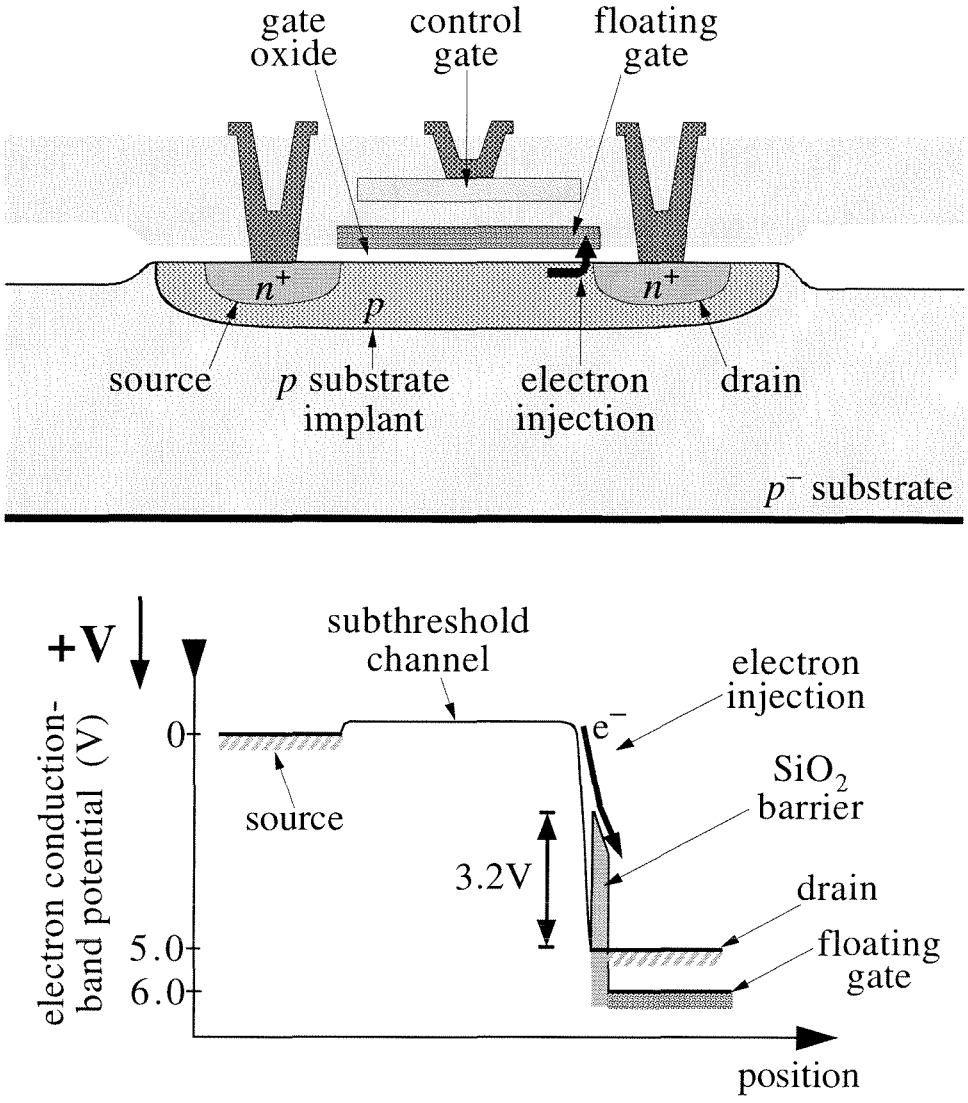


Figure 1.3 Electron injection decreases the weight. I inject electrons from the drain-to-channel depletion region of a subthreshold MOSFET to the floating gate. The *n*FET transistor shown here differs from a conventional *n*-type MOSFET in its use of a moderately doped channel implant. This implant facilitates hot-electron injection, for reasons that I discuss in Section 2.1.2. The floating-gate *p*FET devices, by contrast, induce a hot-electron gate current without any special channel implant; they generate electrons for oxide injection by means of hole-impact ionization in the drain-to-channel depletion region of a subthreshold *p*-type MOSFET (see Section 3.1.2).

1.3.2.3 Weight Updates Are Bidirectional

I effect bidirectional weight updates by using FN tunneling to remove electrons from the floating gate, and by using hot-electron injection to add electrons to the floating gate. The tun-

neling and injection oxide currents vary with the transistor's terminal voltages and source current; consequently, the weight-update rate $\partial W/\partial t$ varies with the terminal voltages, which are imposed on the device, and with the source current, which is the present output. As a result, the synapse learns: Its future weight value depends on the applied input and the present weight value.

1.3.2.4 Floating-Gate Storage Is Nonvolatile

Thermally accelerated leakage experiments on floating-gate MOS devices fabricated in the 2μm Orbit process indicate that the weight will decrease by less than an e-fold in magnitude over a 25-year period at 55°C [36].

1.3.3 Floating-Gate MOS Synapse Transistors

I have named my devices *synapse transistors*. In Chapter 2, I describe and characterize a pair of *n*-type synapse transistors; these devices integrate FN tunneling and hot-electron injection within an *n*-type floating-gate MOSFET. In Chapter 3, I describe and characterize a pair of *p*-type synapse transistors; these devices integrate FN tunneling and hot-electron injection within a *p*-type floating-gate MOSFET. All four devices are compact, consume little power, and operate off a single-polarity supply.

Although a single transistor cannot model the complex behavior of a neural synapse completely, my synapse transistors do possess the nine attributes that I enumerated in Section 1.3, and they can learn from an input signal without interrupting the ongoing computation: Their future output depends on both the applied input and the present output. Because synapse transistors permit both local computation and local weight updates, I can use them to build autonomous learning arrays in which both the system outputs, and the memory updates, are computed locally and in parallel (see Chapter 4).

References

1 C. S. Sherrington, "The central nervous system," in *A Text Book of Physiology,* M. Foster, ed., 7th edition, London: Macmillian, 1897.

2 P. S. Churchland and T. A. Sejnowski, *The Computational Brain,* Cambridge, MA: MIT Press, 1993.

3 L. D. Harmon, "Natural and artificial synapses," in *Self-Organizing Systems 1962,* M. C. Yovits, G. T. Jacobi, and G. D. Goldstein, eds., Washington, D. C.: Spartan, 1962.

4 E. R. Kandel, J. H. Schwartz, and T. M. Jessel, *Principles of Neural Science,* 3rd ed., Norwalk, CT: Appleton & Lange, 1991.

5 R. Douglas, "Rules of thumb for neuronal circuits in the neocortex," *Notes for the Neuromorphic aVLSI Workshop,* Telluride, CO, 1994.

6 T.V.P. Bliss and T. Lømo, "Long-term potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path," *J. of Physiology,* vol. 232, pp. 331–356, 1976.

7 G. Barrionuevo, S. R. Kelso, D. Johnston, and T. H. Brown, "Conductance mechanism responsible for long-term potentiation in monosynaptic and isolated excitatory synaptic inputs to hippocampus," *J. of Neurophysiology,* vol. 55, no, 3, pp. 540–550, 1986.

8 C. Koch, "Computation and the single neuron," *Nature,* vol. 385, no. 6613, pp. 207–210, 1997.

9 C. F. Stevens, "Strengths and weaknesses in memory," *Nature,* vol. 381, no. 6582, pp. 471–472, 1996.

10 H. Markram, J. Lubke, M. Frotscher, and B. Sakmann, "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs," *Science,* vol. 275, no. 5297. pp. 213–215, 1997.

11 S. Ramón y Cajal, *Histology of the Nervous System of Man and Vertebrates,* vol. 2, translated from the French version of the original Spanish by N. Swanson and L. W. Swanson, New York: Oxford University Press, 1995.

12 D. O. Hebb, *The Organization of Behavior*, New York: Wiley, 1949.

13 W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophysics,* vol. 5, pp. 115–133, 1943.

14 F. Rosenblatt, *Principles of Neurodynamics,* New York: Spartan, 1962.

15 J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *P. Natl. Acad. Sci.,* vol. 79, no. 8, pp. 2554–2558, 1982.

16 D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature,* vol. 323, no. 6088, pp. 533–536, 1986.

17 C. Mead, "Neuromorphic electronic systems," *P. of the IEEE,* vol. 78, no. 10, pp. 1629–1636, 1990.

18 M. L. Minsky and S. A. Papert, *Perceptrons,* Cambridge, MA: MIT Press, 1969.

19 K. J. Jeffery and I. C. Reid, "Modifiable neuronal connections: An overview for psychiatrists," *Am. J. of Psychiatry,* vol. 154, no. 2, pp. 156–164, 1997.

20 M. A. Mahowald and R. J. Douglas, "A silicon neuron," *Nature,* vol. 354, no. 6354, pp. 515–518, 1991

21 J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation,* Reading, MA: Addison-Wesley, 1991.

22 B. Hochet, V. Peiris, S. Abdo, and M. J. Declercq, "Implementation of a learning Kohonen neuron based on a new multilevel storage technique," *IEEE J. Solid-State Circuits*, vol. 26, no. 3, pp. 262–267, 1991.

23 C. Schneider and H. Card, "Analog CMOS synaptic learning circuits adapted from invertebrate biology," *IEEE Trans. Circuits and Systems,* vol. 38, no. 12, pp. 1430–1438, 1991.

24 D. Kerns, J. Tanner, M. Sivilotti, and J. Luo, "CMOS UV-writeable nonvolatile analog storage," in *Advanced Research in VLSI,* E. Sequin, ed., Cambridge, MA: MIT Press, pp. 245–261, 1991.

25 M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable artificial neural network (ETANN) with 10240 'floating gate' synapses," *Proc. 1991 Intl. Joint Conf. Neural Networks,* Washington D.C., vol. 2, pp. 191–196, 1989.

26 H. C. Card and W. R. Moore, "Silicon models of associative learning in aplysia," *Neural Networks,* vol. 3, no. 3, pp. 333–346, 1990.

27 H. C. Card, C. Schneider, and W. R. Moore, "Hebbian plasticity in MOS synapses," *IEE Proc. F. RADAR and Sig. Processing,* vol. 138, no. 1, pp. 13–16, 1991.

28 J. Alspector and R. Allen, "A neuromorphic VLSI learning system," in *Proceedings of the 1987 Stanford Conference on Advanced Research in VLSI,* P. Losleben, ed., Cambridge, MA: MIT Press, pp. 313–349, 1987.

29 J. Lazzaro, J. Wawrzynek, and A. Kramer, "Systems technologies for silicon auditory models," *IEEE Micro*, vol. 14, no. 3, pp. 7-15, 1994.

30 D. A. Durfee and F. S. Shoucair, "Comparison of floating-gate neural-network memory cells in standard VLSI CMOS technology," *IEEE Trans. Neural Networks,* vol. 3, no. 3, pp. 348–353, 1992.

31 T. Shibata, H. Kosaka, H. Ishii, and T. Ohmi, "A neuron-MOS neural-network using self-learning compatible synapse circuits," *IEEE J. Solid-State Circuits,* vol. 30, no. 8, pp. 913–922, 1995.

32 F. Masuoka, R. Shirota, and K. Sakui, "Reviews and prospects of nonvolatile semiconductor memories," *IEICE Trans.,* vol. E 74, no. 4, pp. 868–874, 1991.

33 D. Kahng and S. M. Sze, "A floating-gate and its applications to memory devices," *The Bell System Technical Journal,* vol. 40, pp. 1288–1295, July–August, 1967.

34 D. Kahng, "Semipermanent memory using capacitor charge storage and IGFET readout," *The Bell System Technical Journal,* vol. 40, pp. 1296–1300, July–August, 1967.

35 D. Frohman-Bentchkowsky, "Memory behavior in a floating-gate avalanche-injection MOS (FAMOS) structure," *Appl. Phys. Lett.,* vol. 18, no. 8, pp. 332–334, 1971.

36 H. Yang, B. J. Sheu, and J. C Lee, "A nonvolatile analog memory using floating-gate MOS transistors," *Analog Integrated Circuits and Signal Processing,* vol. 2, no. 1, pp. 19–25, 1992.

37 O. Fujita and Y. Amemiya, "A floating-gate analog memory device for neural networks," *IEEE Trans. E.D.,* vol. 40, no. 11, pp. 2029–2035, 1993.

38 C. K. Sin, A. Kramer, V. Hu, R. Chu, and P. Ko, "EEPROM as an analog storage device, with particular applications in neural networks," *IEEE Trans. E.D.,* vol. 39, no. 6, pp. 1410–1419, 1992.

39 C. A. Mead, *Analog VLSI and Neural Systems,* Reading, MA: Addison-Wesley, 1989.

40 M. Lenzlinger and E. H. Snow, "Fowler–Nordheim tunneling into thermally grown SiO_2," *J. of Appl. Phys.*, vol. 40, no. 6, pp. 278–283, 1969.

41 E. Takeda, C. Yang, and A. Miura-Hamada, *Hot-Carrier Effects in MOS Devices,* San Diego, CA: Academic Press, 1995.

42 A. S. Grove, *Physics and Technology of Semiconductor Devices,* New York: John Wiley & Sons, 1967.

Chapter 2

The *n*-Type Synapse Transistors

In this chapter, I describe the layout, characteristics, and weight-update behavior of my *n*FET synapse transistors. I describe first a four-terminal synapse, and from this device I then develop a three-terminal synapse. I describe my *p*-type synapse transistors in Chapter 3.

2.1 A Four-Terminal *n*FET Synapse

The four-terminal *n*FET synapse, shown in Figure 2.1, is an *n*-type MOSFET with a poly1 floating gate, a poly2 control gate, a moderately doped channel, and a fourth terminal used for gate-oxide tunneling. I use FN tunneling to remove electrons from the floating gate, and use channel hot-electron injection (CHEI) to add electrons to the floating gate. This *n*FET synapse has the following features:

- Electrons tunnel from the floating gate to an n^+ tunneling implant through gate oxide. High voltages applied to the tunneling implant provide the oxide electric field required for tunneling. To prevent breakdown of the reverse-biased *pn* junction from the substrate to the tunneling implant, I surround the n^+ tunneling implant with a lightly doped ($\sim 5\times10^{15}\text{cm}^{-3}$) n^- well. Tunneling removes electrons from the floating gate, increasing the synapse weight *W*.

- Electron tunneling is enhanced where the poly1 floating gate overlaps the heavily doped ($\sim 1\times10^{20}\text{cm}^{-3}$) n^+ tunneling implant, for two reasons. First, the gate cannot deplete the n^+, whereas it does deplete the n^- well. Thus, the oxide electric field is higher over the n^+. Second, enhancement at the gate edge further augments the oxide field.

- Electrons inject from the drain-to-channel space-charge region to the floating gate. To facilitate CHEI, I apply a bulk *p*-type implant (a $\sim 1\times10^{17}\text{cm}^{-3}$ NPN BJT base implant) to the

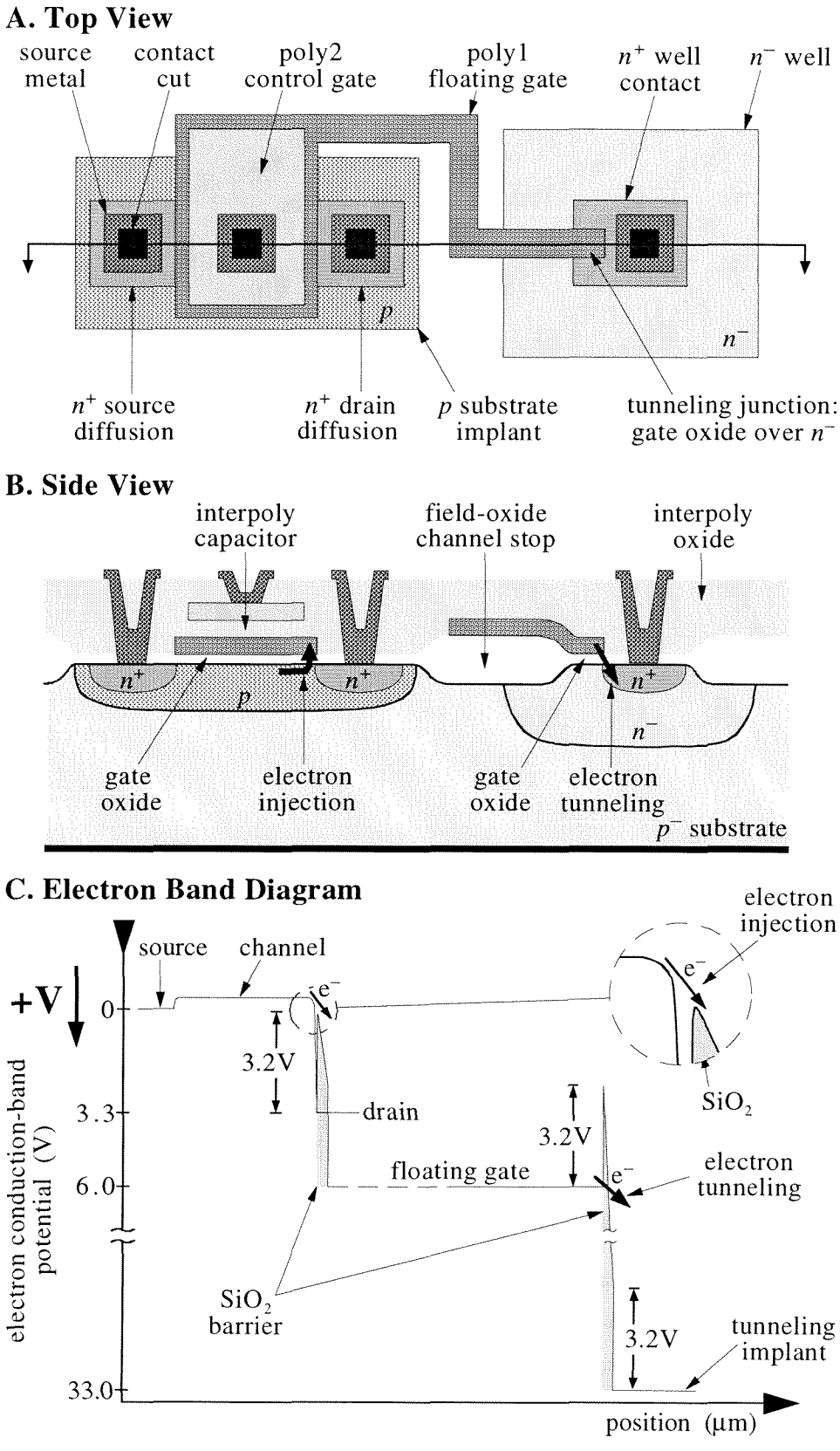


Figure 2.1 The four-terminal *n*FET synapse, showing the electron tunneling and injection locations. The

three diagrams are aligned vertically. Diagrams A and C are drawn to scale; for clarity, I have exaggerated the vertical scale in diagram B. In the 2μm Orbit process, the synapse length is 48μm, and the width is 17μm. All voltages in the conduction-band diagram are referenced to the source potential, and I have assumed subthreshold source currents (I_s<100nA). Although the gate-oxide band diagram actually projects into the plane of the page, for clarity I have rotated it by 90° and have drawn it in the channel direction. When compared with a conventional *n*FET, the *p*-type substrate implant quadruples the MOS gate-to-channel capacitance. With a 50fF interpoly capacitor as shown, the coupling coefficient between the poly2 control gate and the poly1 floating gate is only 0.2. To facilitate testing, I enlarged the interpoly capacitor to 1pF, thereby increasing the coupling to 0.8. This device, like all my synapse transistors, requires large tunneling voltages, because the gate-oxide thickness in the 2μm process ranges from 350–450Å. Consequently, I surround the n^+ tunneling implant with an n^- well, to prevent breakdown of the reverse-biased *pn* junction from the *n*-type tunneling implant to the *p*-type substrate.

MOS-transistor channel. This implant serves two functions. First, it increases the peak drain-to-channel electric field, thereby increasing the hot-electron population in the drain-to-channel depletion region. Second, it allows the MOSFET to operate with both high floating-gate voltages and subthreshold source currents; if the floating-gate voltage exceeds the drain voltage, the drain-to-gate oxide electric field transports injected electrons to the floating gate. CHEI adds electrons to the floating gate, decreasing the synapse weight *W*.

- Oxide uniformity and purity determine the initial matching between synapses, as well as the learning-rate degradations due to oxide trapping. I therefore use the thermally grown gate oxide for all SiO_2 carrier transport.

I intend to build silicon-learning systems using synapse transistors with subthreshold channel currents (note that I equate channel current with source current). The learning behavior of my systems will derive in part from the tunneling and injection processes that alter the stored weights; consequently, I have investigated these processes over the subthreshold channel-current range.

2.1.1 Electron Tunneling Increases the Weight

I increase the synapse weight *W* by tunneling electrons off the floating gate [1]. The tunneling process is shown in the energy-band diagram [2] of Figure 2.1. A potential difference between the tunneling implant and the floating gate reduces the effective oxide thickness, facilitating electron tunneling from the floating gate, through the SiO_2 barrier, into the oxide conduction

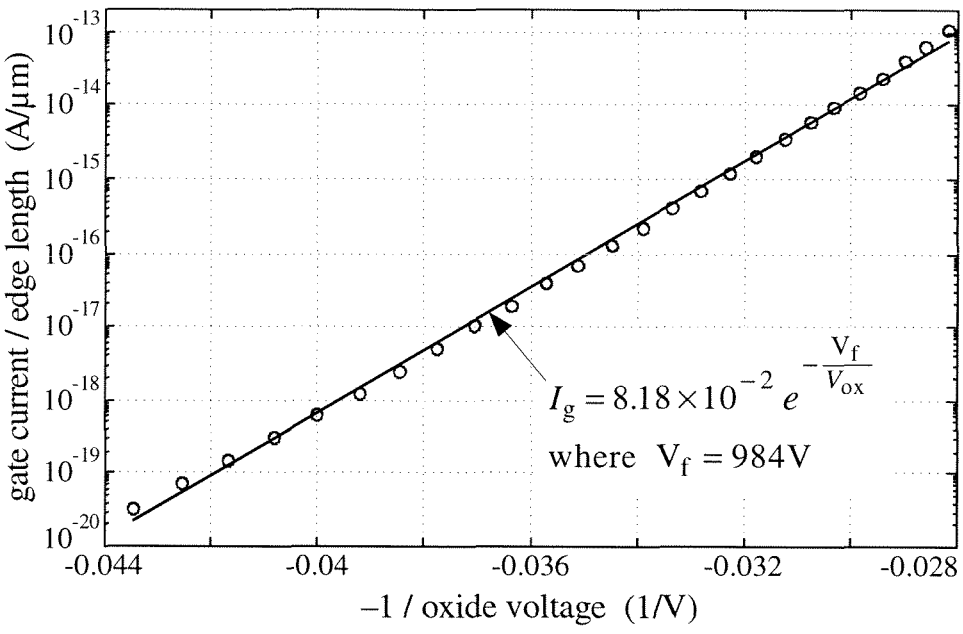


Figure 2.2 Tunneling gate current versus reciprocal oxide voltage. I measured the tunneling gate current I_g versus the oxide voltage V_{ox}, where I define V_{ox} to be the potential difference between the n^+ tunneling implant and the floating gate, and I plotted I_g versus $-1/V_{ox}$. I fit the data using a conventional Fowler–Nordheim expression [1, 3]. I normalized I_g to the tunneling-junction gate-to-n^+ edge length, in lineal microns, because the floating gate induces a depletion region in the lightly doped n^- well, reducing the effective oxide voltage and with it the tunneling current. Because the gate cannot deplete the n^+ well contact appreciably, the oxide electric field is higher where the self-aligned floating gate overlaps the n^+. Because I_g increases exponentially with V_{ox}, gate-oxide tunneling in the synapse transistors is primarily an edge phenomenon.

band. These electrons are then swept over to the tunneling implant by the oxide electric field. I apply positive high voltages to the tunneling implant to promote electron tunneling.

In Figure 2.2, I show the tunneling gate current (the oxide current) versus the reciprocal of the voltage across the tunneling oxide. I fit these data with an FN fit [1, 3]:

$$I_g = \mathrm{I}_{to}\, e^{-\frac{V_f}{V_{ox}}} \tag{2.1}$$

where I_g is the gate current; V_{ox} is the oxide voltage; V_f=984V is consistent with a recent survey [4] of SiO_2 tunneling, given the synapse transistor's 400Å gate oxide; and I_{to} is a pre-exponential current.

2.1.2 Electron Injection Decreases the Weight

I decrease the synapse weight W by injecting electrons onto the floating gate. CHEI in conventional MOSFETs is well known [5]. It occurs in short-channel devices with continuous channel currents, when a high gate voltage is combined with a large potential drop across the short channel. It also occurs in switching transistors, when both the drain and gate voltages are transiently high. In neither case is the CHEI suitable for use in a learning system. The short-channel CHEI requires large channel currents, consuming too much power; the switching-induced CHEI is a poorly controlled transient phenomenon. Instead, I use the drain-to-channel electric field in a subthreshold MOSFET to accelerate channel electrons to high energies; I show the process in the energy-band diagram of Figure 2.1.

Electrons inject from the transistor channel, over the 3.2V Si–SiO_2 work-function barrier, into the oxide conduction band. These electrons then are swept over to the floating gate by the oxide electric field. For electrons to be collected at the floating gate, the following three conditions must be satisfied: (1) the electrons must possess the 3.2eV required to surmount the Si–SiO_2 work-function barrier, (2) the electrons must scatter upward into the gate oxide, and (3) the oxide electric field must be oriented in the proper direction to transport the injected electrons to the floating gate.

In a conventional n-type MOSFET, requirements 1 and 2 are readily satisfied: I merely operate the transistor in its subthreshold regime, with a drain-to-source voltage greater than about 3V. Because the subthreshold channel-conduction band is flat, the drain-to-channel transition is steep, and the drain-to-channel electric field is large. Channel electrons are accelerated rapidly in this field; a fraction of them acquire the 3.2eV required for hot-electron injection. A fraction of these 3.2eV electrons naturally scatter, by means of collisions with the semiconductor lattice, upward into the gate oxide.

It is principally requirement 3 that prevents a gate current in a conventional subthreshold nFET. Subthreshold operation typically implies gate-to-source voltages less than 0.8V. With the drain at 3V, and the gate at 0.8V, the drain-to-gate electric field opposes the transport of the injected electrons to the floating gate. The electrons are instead returned to the drain.

In the synapse transistor, I promote the transport of injected electrons to the floating gate by adding a bulk p-type implant to the channel region. This implant serves two functions. First, it increases the peak drain-to-channel electric field, thereby increasing the hot-electron population in the drain-to-channel depletion region. Second, it increases the channel surface–acceptor concentration, raising the transistor's threshold voltage V_t from 0.8V to 6V. This increase ensures that, for typical floating-gate and drain voltages of about 5.5V and 3V, respectively, the channel

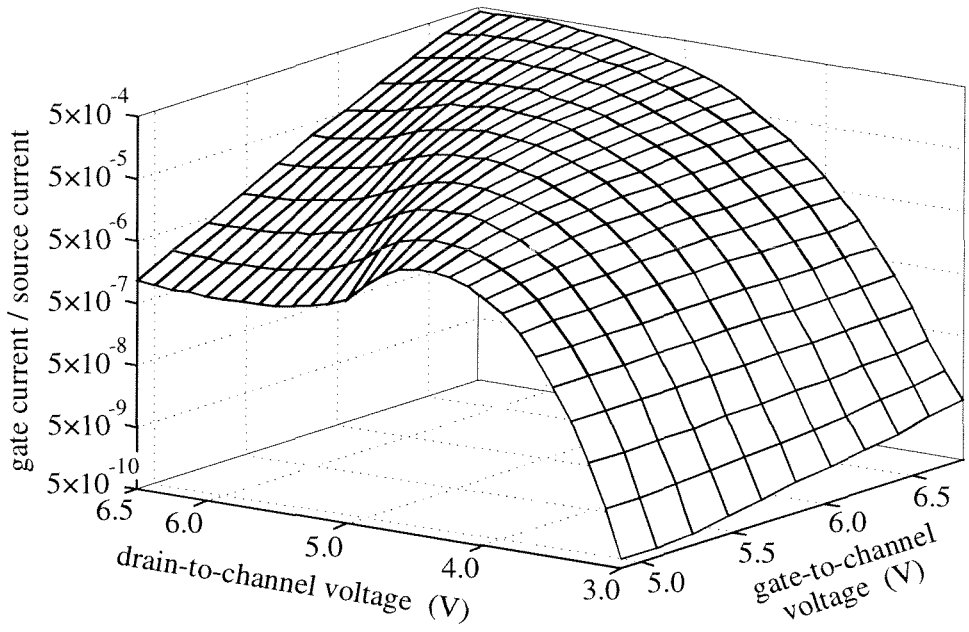


Figure 2.3 Four-terminal *n*FET-synapse CHEI surface plot. I measured the CHEI gate current I_g versus both the gate-to-channel potential, V_{gc}, and the drain-to-channel potential, V_{dc}, for a fixed source current I_s=1 nA. I plotted the gate current I_g divided by the source current I_s. In the subthreshold regime, I_g increases linearly with I_s (see Figure 2.5); consequently, these data show the CHEI efficiency for the entire subthreshold source-current range. Where $V_{fg}>V_d$ (V_{fg} and V_d are the floating-gate and drain voltages, respectively), the CHEI efficiency is only weakly dependent on V_{gc}. Where $V_d>V_{fg}$, the drain-to-gate oxide electric field returns injected electrons to the drain, rather than transporting them to the floating gate, and the CHEI efficiency drops. I anticipate that, for most learning applications, V_{gc} will vary by at most a few hundred millivolts, and V_{fg} always will exceed V_d during CHEI. Consequently, in fit Eqn. (2.2), I assume that the CHEI efficiency depends only on V_{dc}.

current still is subthreshold, but now the drain-to-gate oxide electric field transports injected electrons over to the floating gate, rather than returning them to the drain.

From the perspective of the control gate, raising the MOSFET's threshold voltage is inconsequential, because the control and floating gates are isolated capacitively. I can, therefore, use control-gate inputs in the conventional 0V to 5V range, regardless of the floating-gate transistor's threshold voltage.

In Figure 2.3, I show the four-terminal *n*FET synapse's CHEI efficiency (gate current I_g divided by source current I_s) versus both the drain-to-channel and the gate-to-channel potentials. I plot the data as efficiency because the gate current increases linearly with the source current over the entire subthreshold range (see Figure 2.5). I reference the drain to the channel potential be-

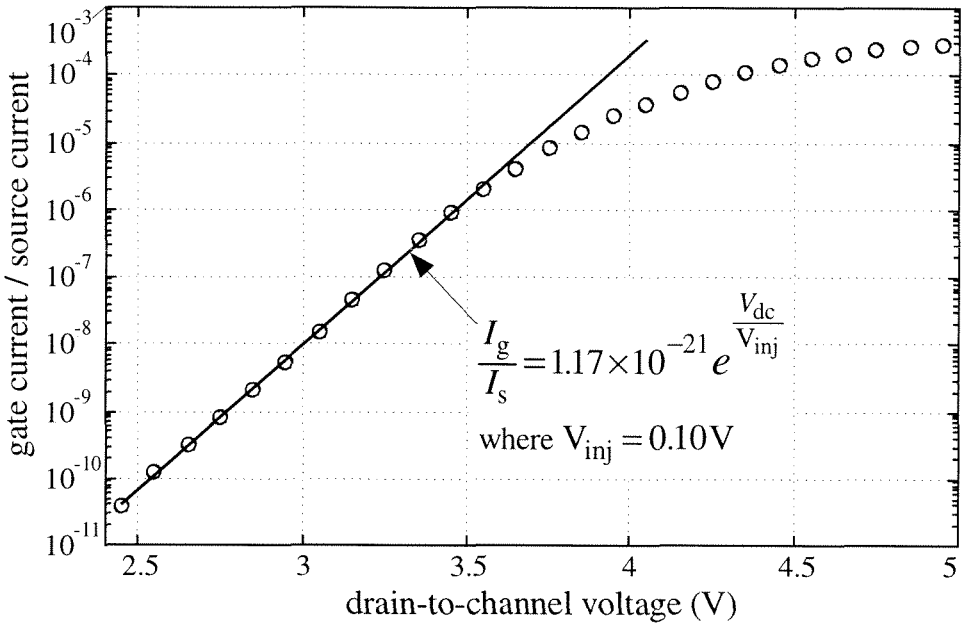


Figure 2.4 Four-terminal *n*FET-synapse CHEI efficiency versus drain-to-channel voltage. I fixed the gate-to-channel voltage at V_{gc}=6.1 V and the source current at I_s=10 nA, and I measured the gate current I_g versus the drain-to-channel voltage V_{dc}. I anticipate that most learning systems will employ small gate currents; consequently, the simple exponential fit models the data accurately for the range of drain voltages that I expect to use in my learning systems.

cause the hot-electron population derives from the drain-to-channel electric field; I reference the floating gate to the channel potential because the direction of electron transport within the oxide derives from the direction of the gate-to-channel electric field. I can re-reference my results to the source potential by using the relationship between source and channel potential in a sub-threshold MOSFET [6, 7].

In Figure 2.4, I plot both the measured CHEI efficiency, and an empirical fit to these data, versus the drain-to-channel potential, V_{dc}, for a typical value of gate-to-channel potential. When V_{dc} is less than 2 V, the CHEI gate current is exceedingly small, and the weight W remains non-volatile. When V_{dc} exceeds 2.5 V, the CHEI gate current causes measurable changes in the synapse weight W. I expect my silicon-learning systems to use slow adaptation; consequently, I anticipate using very small oxide currents. Because the four-terminal *n*FET synapse's CHEI efficiency is high, I anticipate that V_{dc} typically will be less than 3 V, and always will be less than 3.5 V. Consequently, I can fit the data of Figure 2.4 using a simple exponential:

$$I_g = \beta I_s e^{\frac{V_{dc}}{V_{inj}}} \tag{2.2}$$

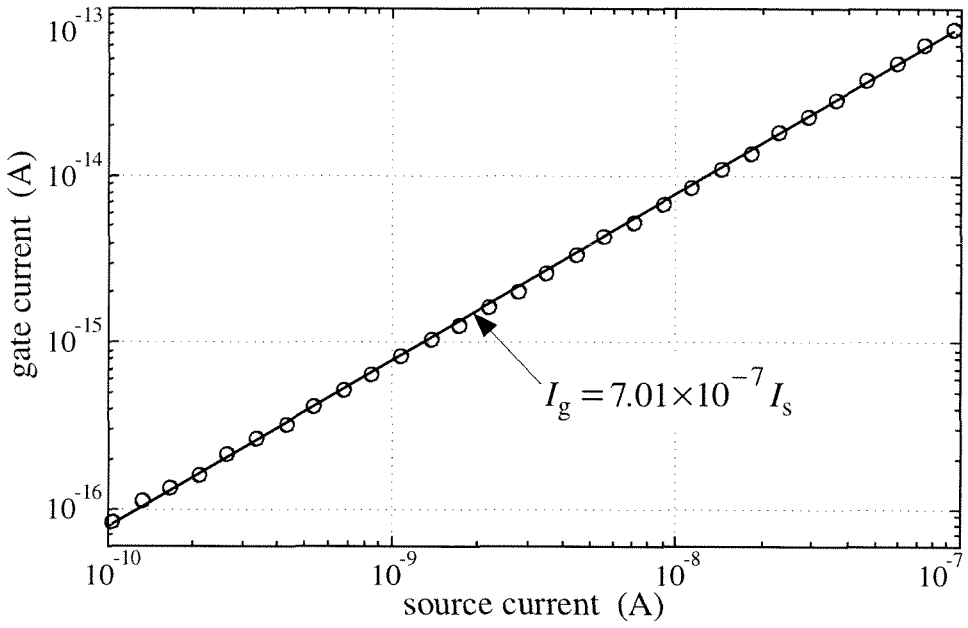


Figure 2.5 Four-terminal *n*FET-synapse gate current versus source current. I held the drain-to-bulk and gate-to-bulk voltages fixed at V_{db}=5V and V_{gb}=7V, respectively, and measured the gate current I_g versus the source current I_s. These data show that the four-terminal *n*FET synapse's CHEI efficiency is independent of source current over the subthreshold source-current range.

where I_g is the gate current; I_s is the source current; V_{dc} is the drain-to-channel potential; and β, V_{inj} are fit constants.

As a consequence of this synapse transistor's 6V threshold, the floating-gate voltage usually exceeds 5V; if V_{dc}<3.5V, then the drain-to-gate oxide electric field strongly favors the transport of injected electrons to the floating gate. The CHEI efficiency therefore is, to first order, independent of the gate-to-channel potential, and I model the CHEI process using only Eqn. (2.2).

2.1.3 The Gate-Current Equation

Because the tunneling and CHEI gate currents flow in opposite directions, I obtain the four-terminal *n*FET synapse's gate-current equation by subtracting Eqn. (2.2) from Eqn. (2.1):

$$I_g = I_{to} e^{-\frac{V_f}{V_{ox}}} - \beta I_s e^{\frac{V_{dc}}{V_{inj}}} \tag{2.3}$$

This equation describes the gate current accurately over my anticipated synapse-operating range.

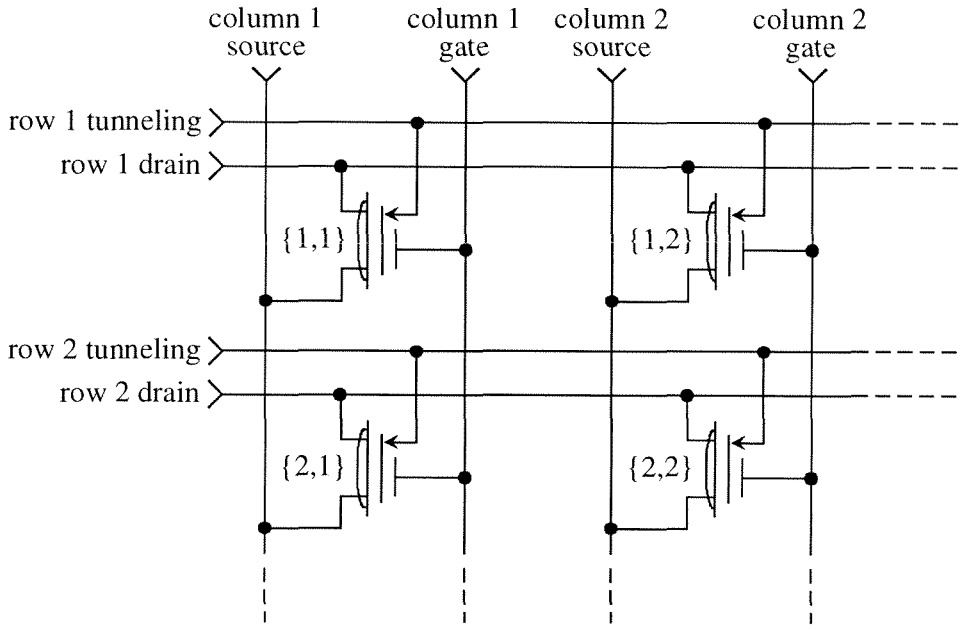


Figure 2.6 A 2×2 array of four-terminal *n*FET synapses. The arrow at each synapse's floating gate denotes a tunneling junction; the curved line in the transistor symbol denotes a pbase channel implant. The row synapses share common tunneling and drain wires; consequently, tunneling or injection at one row synapse can cause undesired tunneling or injection at other row synapses.

2.1.4 Isolation and Weight Updates in a Synaptic Array

A synaptic array, with a synapse transistor at each node, can form the basis of a silicon learning system. I fabricated a simplified 2×2 array of four-terminal *n*FET synapses to investigate isolation during tunneling and injection, and to measure the synapse weight-update rates. Because this 2×2 array uses the same row–column addressing that I will employ in larger arrays, it allows me to characterize the synapse isolation and weight-update rules completely.

I show the array in Figure 2.6. I chose, from among the many possible ways of using the array, to select source current as the synapse output, and to turn off the synapses during tunneling. I applied the voltages shown in Table 2.1 to read, tunnel, or inject synapse {1,1} selectively, while ideally leaving the other synapses unchanged.

2.1.4.1 Synapse Isolation

The tunneling and drain terminals of the array synapse transistors connect within rows, but not within columns. Consequently, the tunneling and CHEI crosstalk between column synapses is negligible. A synapse's tunneling gate current increases exponentially with its oxide voltage V_{ox},

Table 2.1 Four-terminal *n*FET-synapse array terminal voltages. I applied these voltages to the array of Figure 2.6, to obtain the data in Figure 2.7.

	column 1 gate	column 2 gate	column 1 source	column 2 source	row 1 drain	row 2 drain	row 1 tunnel	row 2 tunnel
read	+5	0	0	0	+1	0	0	0
tunnel	0	+5	0	0	0	0	+31	0
inject	+5	0	0	0	3.15	0	0	0

(V_{ox}, in turn, decreases linearly with V_{fg}), and its CHEI gate current increases linearly with its channel current I_s, (I_s, in turn, increases exponentially with V_{fg}). Consequently, the tunneling and CHEI crosstalk between row synapses decrease exponentially with the voltage differential between the row synapses' floating gates. By using 5V control-gate inputs, I achieve about a 4V differential between the floating gates of the selected and deselected synapses; the resulting crosstalk between row synapses is <0.01 % for all operations.

I show synapse-isolation data in Figure 2.7. To obtain the data in part A, I first initialized all four synapses to I_s=100pA. I then tunneled the {1,1} synapse up to 100nA, and injected it back down to 100pA, while I measured the source currents of the other three synapses. As I expected, the row 2 synapses were unaffected by either the tunneling or the injection. Coupling to the {1,2} synapse also was small.

To obtain the data in part B of Figure 2.7, I first initialized all four synapses to I_s=100nA. I then injected the {1,1} synapse down to 100pA, and tunneled it back up to 100nA. As in the experiment of part A, crosstalk to the other synapses was negligible.

2.1.4.2 Synapse Weight Updates

A synapse's weight updates derive from the tunneling and CHEI oxide currents that alter the floating-gate charge. Because these oxide currents vary with the synapse's terminal voltages and source current, the weight-update rate $\partial W/\partial t$ varies with the terminal voltages, which are imposed on the device, and with the source current, which is the synapse output. Consequently, the synapse learns: Its future output depends on both the applied input and the present output.

I repeated the experiment of Figure 2.7 (A), for several tunneling and injection voltages; in Figure 2.8, I plot the magnitude of the temporal derivative of the source current versus the source current, for a synapse transistor with (part A) a set of fixed tunneling voltages, and (part B) a set

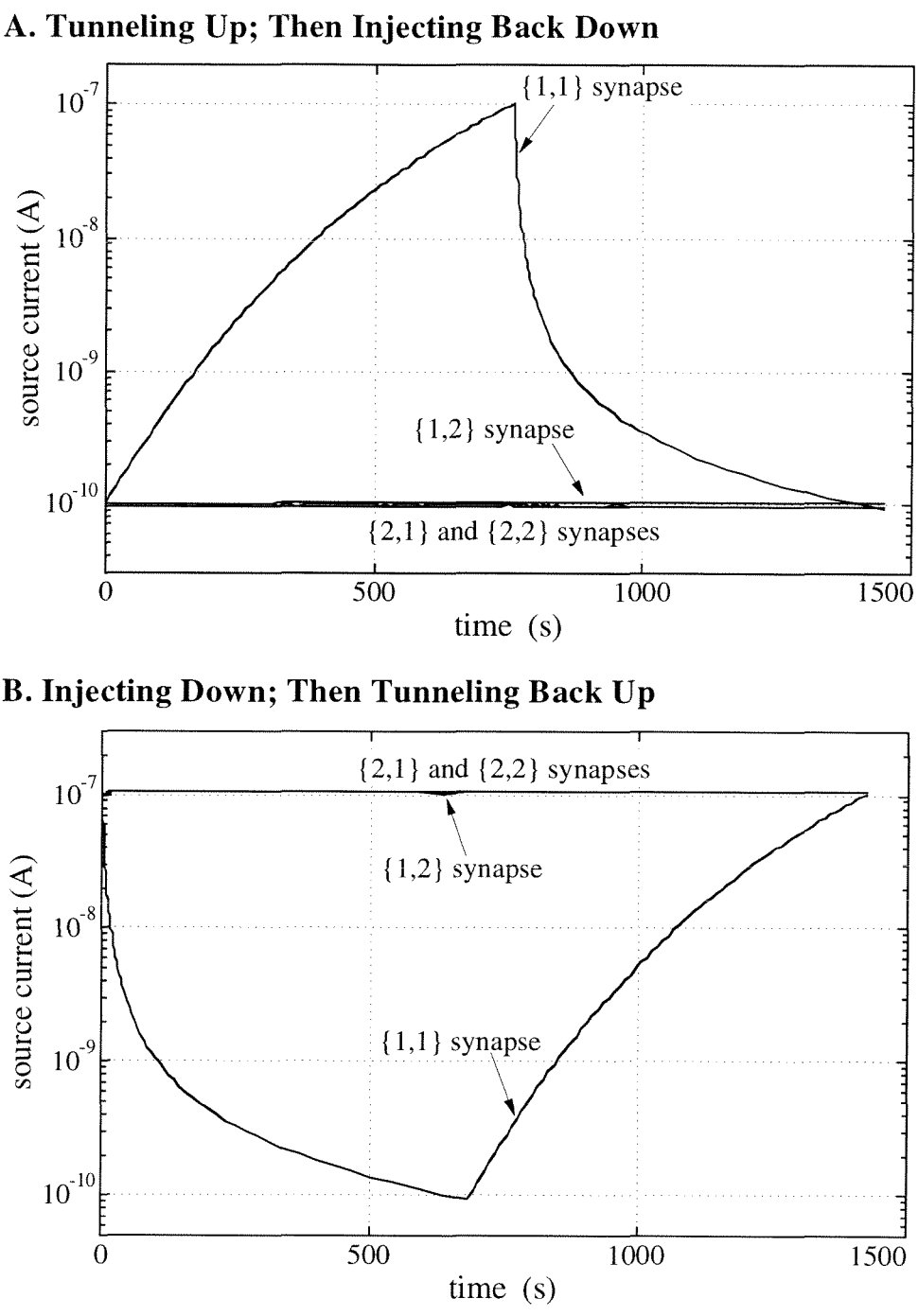


Figure 2.7 Isolation in a 2×2 array of four-terminal *n*FET synapses. The terminal voltages for both experiments are shown in Table 2.1 (see pg. 24). (A) I first tunneled the {1,1} synapse up to 100nA, then I injected it back down to 100pA, while I measured the source currents of the other three synapses. Crosstalk to the {1,2} synapse, defined as the fractional change in the {1.2} synapse's source current divided by the fractional change in the {1,1} synapse's source current, was 0.006% during tunneling, and was 0.002% during injection. (B) I first injected the {1,1} synapse down to 100pA, then I tunneled it back up to 100nA. Crosstalk to the {1,2} synapse was 0.001% during injection, and was 0.002% during tunneling.

of fixed drain voltages. In both experiments, I held the control-gate input V_{in} fixed; consequently, these data show the synapse weight updates $\partial W/\partial t$, as can be seen by differentiating Eqn. (1.2). I now derive a weight-update rule that fits these data.

2.1.4.2.1 The Tunneling Weight-Increment Rule

I first show that tunneling-induced weight increments follow a power law. I begin by taking the temporal derivative of the synapse weight W, where $W \equiv \exp(Q_{fg}/Q_T)$:

$$\frac{\partial W}{\partial t} = \frac{W}{Q_T}\frac{\partial Q_{fg}}{\partial t} = \frac{W}{Q_T} I_g \tag{2.4}$$

I substitute Eqn. (2.1) for the gate current I_g:

$$\frac{\partial W}{\partial t} = \frac{I_{to} W}{Q_T} e^{-\frac{V_f}{V_{ox}}} \tag{2.5}$$

I substitute $V_{ox}=V_{tun}-V_{fg}$ (where V_{tun} and V_{fg} are the tunneling-node and floating-gate voltages, respectively), assume that $V_{tun} \gg V_{fg}$, expand the exponent using $(1-x)^{-1} \approx 1+x$, and solve:

$$\frac{\partial W}{\partial t} \approx \frac{I_{to} W}{Q_T} e^{-\frac{V_f}{V_{tun}} - \frac{V_f V_{fg}}{V_{tun}^2}} \tag{2.6}$$

I substitute $V_{fg}=U_t Q_{fg}/\kappa Q_T$, and solve for the tunneling weight-increment rule:

$$\frac{\partial W}{\partial t} \approx \frac{1}{\tau_{tun}} W^{(1-\sigma)} \tag{2.7}$$

where

$$\sigma \equiv \frac{V_f U_t}{\kappa V_{tun}^2} \tag{2.8}$$

and

$$\tau_{tun} \equiv \frac{Q_T}{I_{to}} e^{\frac{V_f}{V_{tun}}} \tag{2.9}$$

The parameters σ and τ_{tun} vary with the tunneling-node voltage V_{tun}.

2.1.4.2.2 The CHEI Weight-Decrement Rule

I now show that the CHEI-induced weight decrements also follow a power law. I begin by defining a synapse transistor's drain-to-channel potential, V_{dc}, in terms of V_{ds} and I_s. In a subthreshold floating-gate MOSFET, the source current is related to the floating-gate and source voltages [8] by

$$I_s = I_o e^{\frac{\kappa V_{fg} - V_s}{U_t}} \tag{2.10}$$

and the channel-surface potential, Ψ, is related to the floating-gate voltage, V_{fg}, [6, 7] by

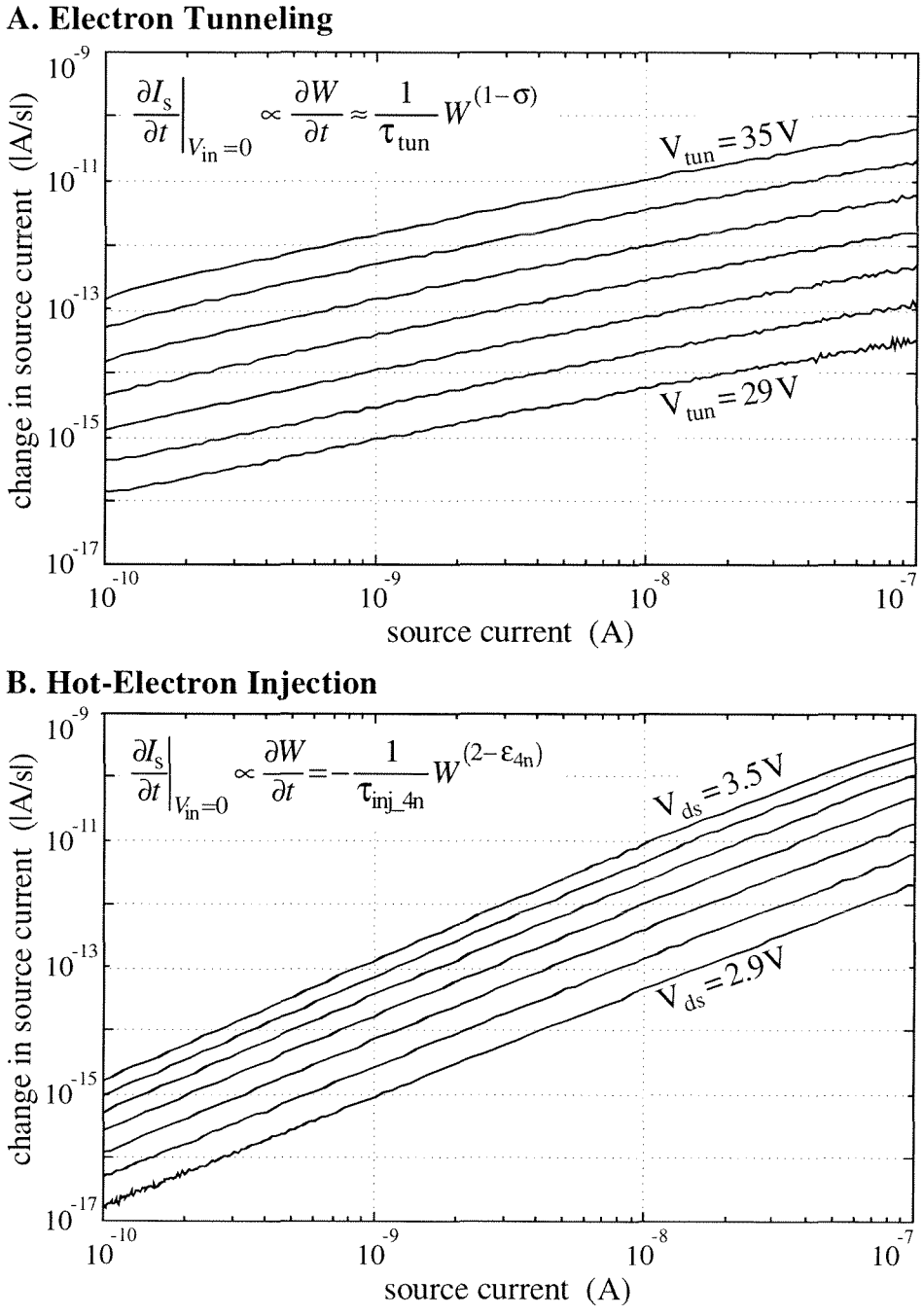


Figure 2.8 Four-terminal *n*FET-synapse (A) tunneling and (B) CHEI weight updates. In both experiments, I measured the synapse's source current I_s versus time, and plotted $|\partial I_s/\partial t|$ versus I_s. I fixed the synapse's terminal voltages; consequently, the change in I_s is a result of changes in the synapse's weight W. In part A, I applied V_{in}=0V, V_s=0V, V_{ds}=2V, and stepped V_{tun} from 29V to 35V in 1V increments; in part B, I applied V_{in}=5V, V_s=0V, V_{tun}=20V, and stepped V_{ds} from 2.9V to 3.5V in 0.1V increments. I turned off the tunneling and CHEI at regular intervals, to measure I_s. Because, for a fixed V_{in}, the synapse weight updates $\partial W/\partial t$ are proportional to $\partial I_s/\partial t$ (see Eqn. (1.2)), these data show that the weight updates follow a power law. The mean values of σ and ε_{4n} are 0.17 and 0.24, respectively.

$$\Psi \approx \kappa V_{fg} + \Psi_o \tag{2.11}$$

where κ is the coupling coefficient from the floating gate to the channel, and Ψ_o derives from the MOS process parameters.

Using Eqns. (2.10) and (2.11), I solve for the surface potential Ψ in terms of I_s and V_s:

$$\Psi = V_s + \Psi_o + U_t \ln\left(\frac{I_s}{I_o}\right) \tag{2.12}$$

I now solve for V_{dc}:

$$V_{dc} = V_d - \Psi = V_{ds} - \Psi_o - U_t \ln\left(\frac{I_s}{I_o}\right) \tag{2.13}$$

The CHEI gate current I_g is given by Eqn. (2.2). I add a minus sign to I_g, because CHEI decreases the floating-gate charge, and substitute for V_{dc} using Eqn. (2.13):

$$I_g = -\beta I_s e^{\frac{V_{ds} - \Psi_o - U_t \ln(I_s/I_o)}{V_{inj}}} = -\beta I_o e^{\frac{U_t}{V_{inj}}} e^{\frac{V_{ds} - \Psi_o}{V_{inj}}} I_s^{\left(1 - \frac{U_t}{V_{inj}}\right)} \tag{2.14}$$

I substitute for I_s using Eqn. (1.2), and solve:

$$I_g = -\beta I_o e^{\frac{\kappa' V_{in}}{U_t} + \frac{V_{ds} - \kappa' V_{in} - \Psi_o}{V_{inj}}} W^{\left(1 - \frac{U_t}{V_{inj}}\right)} \tag{2.15}$$

I substitute Eqn. (2.15) into $\partial W/\partial t$ from Eqn. (2.4),

$$\frac{\partial W}{\partial t} = -\frac{\beta I_o}{Q_T} e^{\frac{\kappa' V_{in}}{U_t} + \frac{V_{ds} - \kappa' V_{in} - \Psi_o}{V_{inj}}} W^{\left(2 - \frac{U_t}{V_{inj}}\right)} \tag{2.16}$$

to get the final weight-decrement rule:

$$\frac{\partial W}{\partial t} = -\frac{1}{\tau_{inj\_4n}} W^{(2 - \varepsilon_{4n})} \tag{2.17}$$

where

$$\varepsilon_{4n} \equiv \frac{U_t}{V_{inj}} \tag{2.18}$$

and

$$\tau_{inj\_4n} \equiv \frac{Q_T}{\beta I_o} e^{-\frac{\kappa' V_{in}}{U_t} - \frac{V_{ds} - \kappa' V_{in} - \Psi_o}{V_{inj}}} \tag{2.19}$$

The parameter ε_{4n} is fixed; the parameter τ_{inj_4n} varies with V_{ds} and with V_{in}.

2.1.4.2.3 The Synapse Weight-Update Rule

I obtain the complete weight-update rule, for the four-terminal *n*FET synapse, by adding Eqns. (2.7) and (2.17):

$$\frac{\partial W}{\partial t} \approx \frac{1}{\tau_{tun}} W^{(1-\sigma)} - \frac{1}{\tau_{inj\_4n}} W^{(2 - \varepsilon_{4n})} \tag{2.20}$$

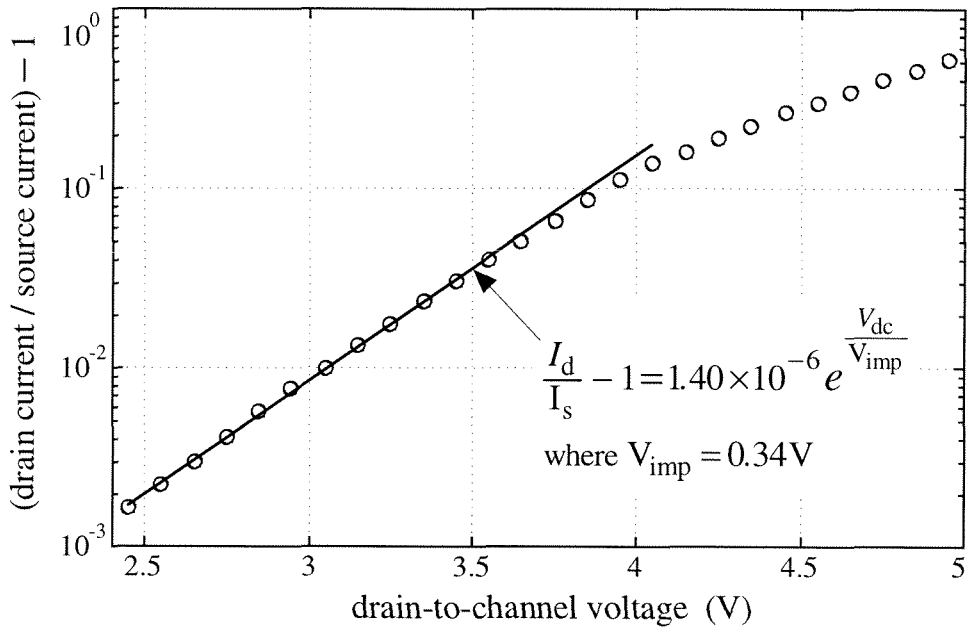


Figure 2.9 Four-terminal *n*FET-synapse impact ionization versus drain-to-channel voltage. I fixed the gate-to-channel voltage at V_{gc}=6.1 V and the source current at I_s=10 nA, and measured the drain current I_d versus the drain-to-channel voltage V_{dc}. By plotting the data as efficiency (drain current I_d divided by source current I_s, minus one), I show the impact-ionization probability as a function of the drain-to-channel voltage. Although I can fit these data over the entire drain-voltage range by using a modified lucky-electron model (see, for example, Figure 2.19), my chosen exponential fit is simpler, and models the data accurately over the drain-voltage range that I anticipate using in my learning systems (I use the same fit range for the CHEI efficiency data in Figure 2.4).

2.1.5 Impact Ionization Increases the Drain Current

In silicon, the barrier energy opposing electron injection into the gate oxide is larger than the activation energy for electron impact ionization; consequently, a drain-to-channel electric field that generates electrons for oxide injection also liberates additional electron–hole pairs [9, 10], causing I_d to increase exponentially with V_{dc}. I show electron impact-ionization data from the four-terminal *n*FET synapse in Figure 2.9. For simplicity, I chose to fit these data using an exponential fit, rather than using the conventional (but more complicated) lucky-electron fit; the exponential fit models the data accurately over the drain-voltage range that I anticipate using in my learning systems. If I choose drain current, rather than source current, as the synapse output, I can rewrite the gate-current equation by replacing I_d with I_s according to my fit equation:

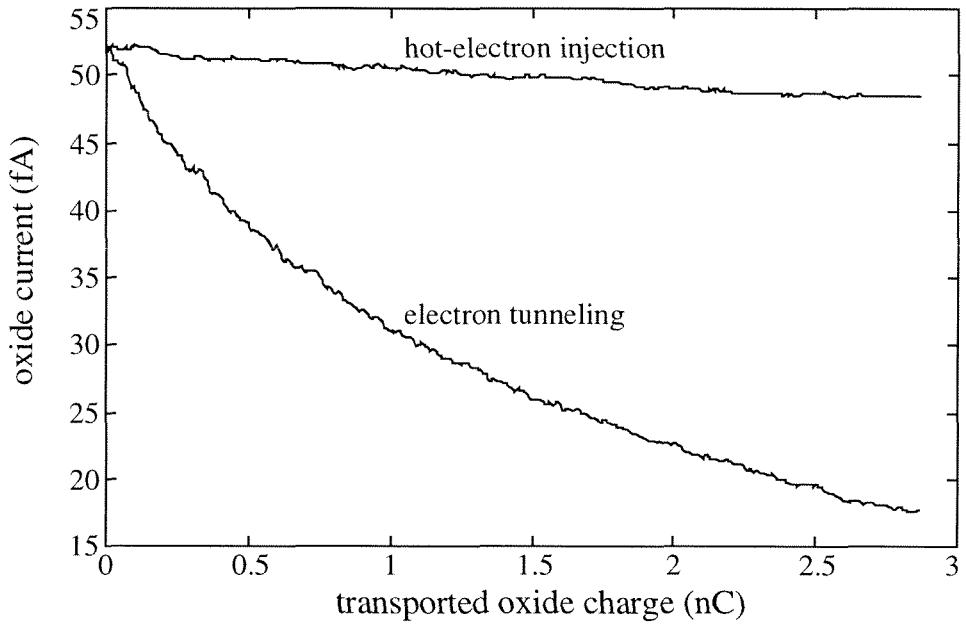


Figure 2.10 Oxide trapping in the four-terminal *n*FET synapse. In the tunneling experiment, I fixed the tunneling-oxide voltage at V_{ox}=31 V. In the injection experiment, I fixed the source current at I_s=50 nA, the drain-to-channel voltage at V_{dc}=3.5 V, and the floating-gate–to–channel voltage at V_{gc}=6.0 V. I measured the tunneling and CHEI oxide currents versus the total (accumulated) charge transported through the oxide. The decrease in oxide current with oxide charge is a consequence of electron trapping. These data show that, to equalize the degradation rates, I should construct future synapses with larger tunneling junctions. Unfortunately, larger tunneling junctions have larger capacitive coupling from the tunneling implant to the floating gate (see the discussion of parasitic coupling in Section 2.3.3). Because the synapse's weight W increases exponentially with the floating-gate charge Q_{fg} (see Eqn. (1.2)), adding even 0.1 nC of charge to the floating gate causes an enormous weight increase; consequently, oxide trapping in the present tunneling junction can be ignored safely.

$$I_d \approx I_s\left(1+\gamma e^{\frac{V_{dc}}{V_{imp}}}\right) \tag{2.21}$$

where γ and V_{imp} are measurable fit constants.

2.1.6 Oxide Trapping Is Small

SiO_2 trapping is a well-known issue in floating-gate transistor reliability [11]. In digital EEPROMs, it ultimately limits the transistor life. In the synapse, oxide trapping decreases the weight-update rate. However, because a synapse transistor's weight W is exponential in the floating-gate charge Q_{fg} (see Eqn. (1.2)), the synapses in a subthreshold-MOS learning system

will transport only small quantities of total oxide charge over the system lifetime. In Figure 2.10, I plot tunneling and CHEI gate currents versus the total (accumulated) charge transported through the oxide. These data show that the CHEI-oxide trapping is small, but that the tunneling-oxide trapping can decrease the tunneling-gate current substantially. To equalize the degradation, I can use larger tunneling junctions. However, I believe that, even without enlarging the junction, the tunneling-oxide trapping can be ignored safely, because synapses require such small quantities of charge for their weight updates.

2.2 A Three-Terminal *n*FET Synapse

The four-terminal *n*FET synapse employs an n^+ doped drain, and an n^+ doped tunneling implant that is surrounded with n^- to prevent *pn* junction breakdown. I now combine these terminals to yield a more compact device. My three-terminal *n*FET synapse integrates the tunneling function within the drain, eliminating the separate tunneling terminal.

The three-terminal *n*FET synapse, shown in Figure 2.11, is an *n*-type MOSFET with a poly1 floating gate, a poly2 control gate, a moderately doped channel, and a lightly doped drain. Like the four-terminal *n*FET synapse, this three-terminal device uses FN tunneling to remove electrons from the floating gate, and CHEI to add them. This synapse's principal features are:

- Electrons tunnel from the floating gate to the drain through gate oxide. High voltages applied to the drain provide the oxide electric field required for tunneling. To prevent breakdown of the reverse-biased *pn* junction from substrate to drain, I use a lightly doped ($\sim 5\times10^{15}\text{cm}^{-3}$) n^- well as the drain (a *well–drain*). Tunneling removes electrons from the floating gate, increasing the synapse weight *W*.

- Electron tunneling is enhanced where the poly1 floating gate overlaps the heavily doped ($\sim 1\times10^{20}\text{cm}^{-3}$) n^+ well–drain contact, for the same two reasons that it is enhanced in the four-terminal *n*FET synapse: (1) the oxide electric field is higher over the n^+, and (2) enhancement at the gate edge further augments the field.

- Electrons inject from the drain-to-channel space-charge region to the floating gate. As I did in the four-terminal *n*FET synapse, I apply a bulk *p*-type ($\sim 1\times10^{17}\text{cm}^{-3}$ NPN BJT base) implant to the three-terminal *n*FET synapse's channel region. This implant serves the same two functions that it does in the four-terminal *n*FET synapse: (1) it increases the peak drain-to-channel electric field, and (2) it allows the MOSFET to operate with much higher floating-gate voltages while retaining subthreshold source currents. Injection adds electrons to the floating gate, decreasing the synapse weight *W*.

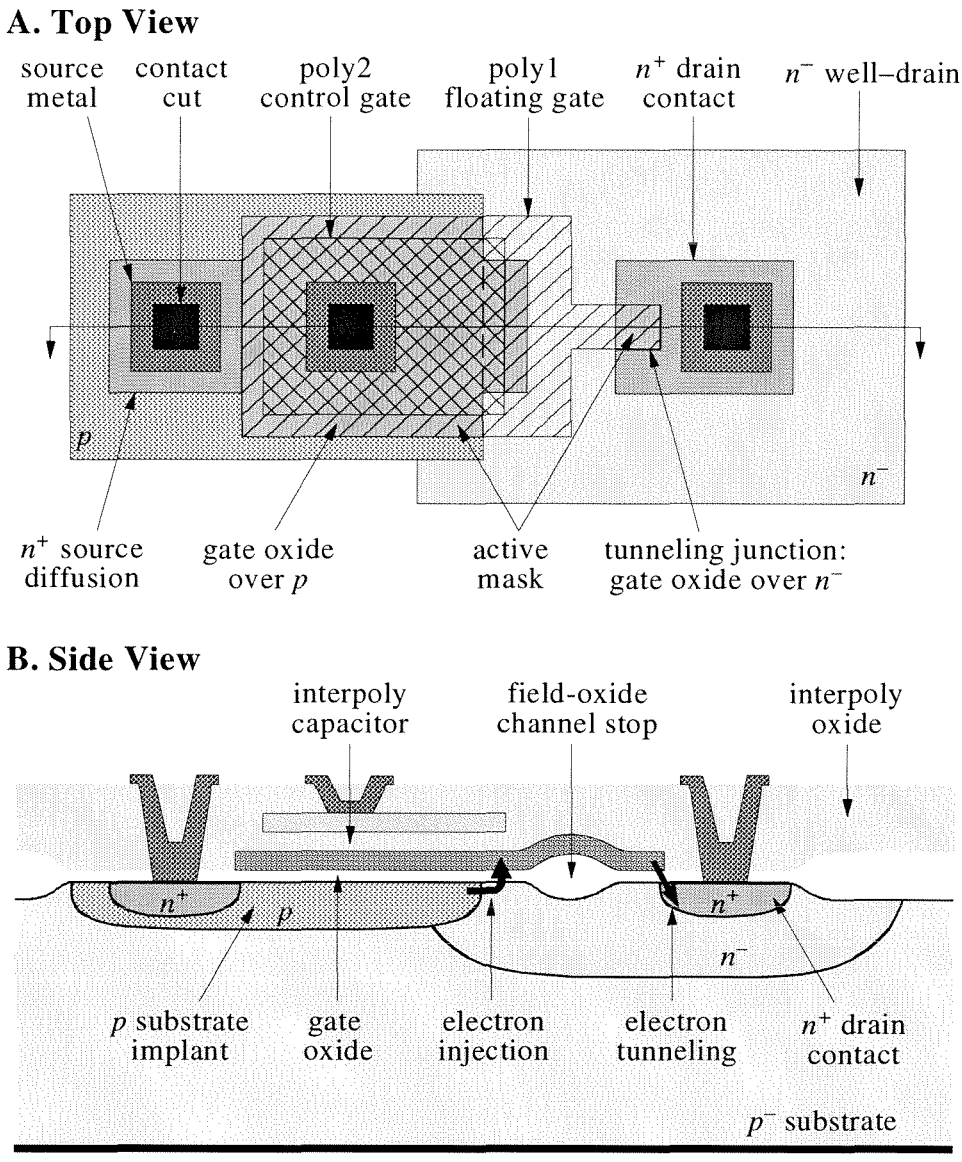


Figure 2.11 The three-terminal *n*FET synapse, showing the electron tunneling and injection locations. The diagrams are aligned vertically. Diagram A is drawn to scale; for clarity, I have exaggerated the vertical scale in diagram B. In the 2μm Orbit process, the synapse length is 38μm, and the width is 16μm. When compared with a conventional *n*FET, the *p*-type substrate implant quadruples the MOS gate-to-channel capacitance. With a 50fF interpoly capacitor as shown, the coupling coefficient between the poly2 control gate and the poly1 floating gate is only 0.2. To facilitate testing, I enlarged the interpoly capacitor to 1pF, thereby increasing the coupling to 0.8.

- Because the channel doping exceeds the drain doping, the drain-to-channel space-charge region appears primarily on the drain side of the channel–drain junction. As a result, the hot-electron population also appears primarily on the drain side of the junction. I greatly facilitate CHEI by extending the MOS gate oxide 2μm beyond the channel–drain edge, over this space-charge region, thereby causing the injected electrons to encounter gate oxide rather than a field-oxide channel stop.

- I use the thermally grown gate oxide for all SiO_2 carrier transport.

I operate the three-terminal *n*FET synapse in or near the subthreshold regime, and I select source current as the output. Because the drain comprises n^- doped rather than n^+ doped silicon, the drain resistance is much higher than in the four-terminal *n*FET synapse. Fortunately, because the channel currents that I use rarely exceed a few microamps; the potential drop within the drain is small; the additional resistance does not affect the transistor operation significantly.

2.2.1 Electron Tunneling Increases the Weight

The tunneling junction in the three-terminal *n*FET synapse is functionally identical to the tunneling junction in the four-terminal *n*FET synapse; consequently, electron tunneling increases the weight W in an identical fashion to that described in Section 2.1.1.

2.2.2 Electron Injection Decreases the Weight

I decrease the three-terminal *n*FET synapse's weight W by injecting electrons onto the floating gate. I show the CHEI process in the energy-band diagram of Figure 2.12 (part A). Like in the four-terminal *n*FET, electrons inject from the transistor channel, over the 3.2V Si–SiO_2 work-function barrier, into the oxide conduction band. These electrons then are swept over to the floating gate by the oxide electric field. For electrons to be collected at the floating gate, the three conditions I enumerated in Section 2.1.2 must again be satisfied: (1) the electrons must possess the 3.2eV required to surmount the Si–SiO_2 work-function barrier, (2) the electrons must scatter upward into the gate oxide, and (3) the oxide electric field must be oriented in the proper direction to transport the injected electrons to the floating gate.

I accelerate channel electrons to 3.2eV in the three-terminal synapse's drain-to-channel electric field. However, a 3.2eV electron population is not, by itself, sufficient for CHEI. As I show in part B of Figure 2.12, a conventional well–drain MOSFET can experience a drain-to-channel electric field greater than 10V/μm, thereby inducing a large 3.2eV electron population. Still, when operating in the subthreshold regime, this device experiences little or no CHEI. Under

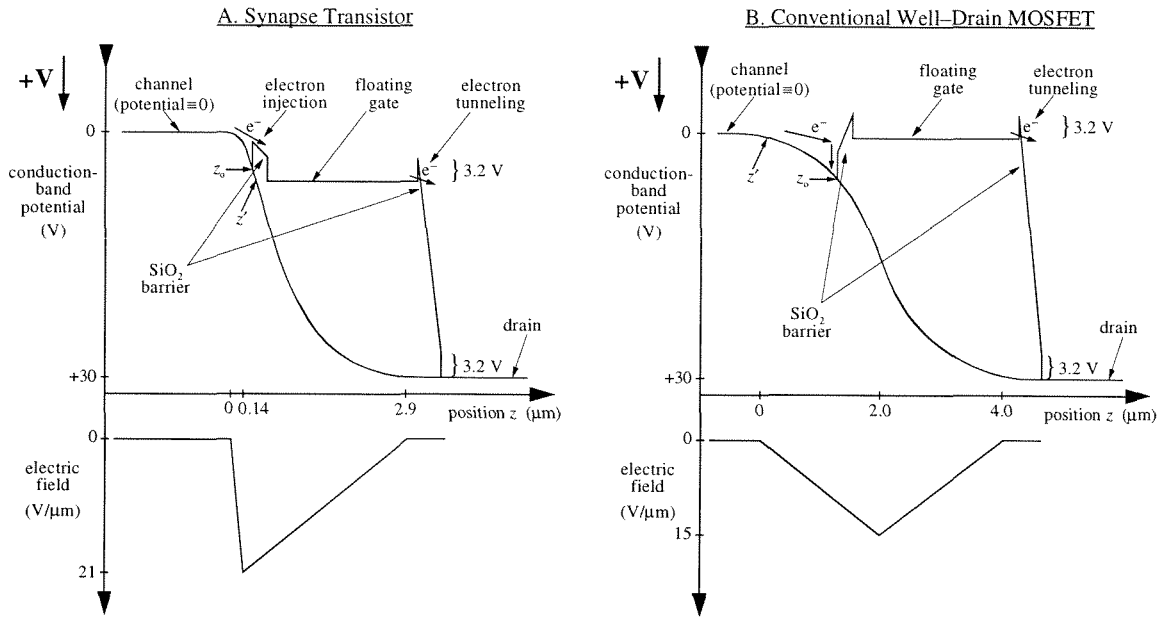


Figure 2.12 The three-terminal *n*FET synapse compared to a well–drain MOSFET. I show, both for the three-terminal *n*FET synapse and for a well–drain MOSFET, the electron conduction-band potential and the corresponding electric field, calculated from implant impurity concentrations [2]. I assume that the well–drain MOSFET is identical to the synapse transistor of Figure 2.11, with the exception of its channel-impurity concentration, which in the synapse transistor is $\sim 1\times10^{17}\text{cm}^{-3}$, and in the well–drain MOSFET is $\sim 5\times10^{15}\text{cm}^{-3}$. For both transistors, I assume step-doping profiles and subthreshold source currents ($I_s<100\text{nA}$). For the synapse transistor, both the impact-ionization data of Figure 2.19, and an observed 70V drain-avalanche onset, are consistent with my step-junction assumption. I reference all voltages to the channel potential; I measure all positions from the channel edge of the drain-to-channel space-charge layer. Although the gate-oxide band diagrams actually project into the plane of the page, for clarity I have rotated them by 90° and have drawn them in the channel direction. Because, for both devices, the conduction-band edge provides the reference potential for the oxide barrier's leading edge, the barrier shape varies with position z along the channel. For clarity, I have drawn the oxide barriers for only a single channel position, z_0. At $z=z'$, the oxide voltage is zero; for $z>z'$, the oxide electric field opposes the transport of injected electrons to the floating gate.

similar conditions, the three-terminal *n*FET synapse's CHEI efficiency can exceed 1×10^{-8}. This improvement is a consequence of the synapse transistor's additional *p*-type channel doping, for two reasons.

First, as a result of the higher channel-impurity concentration, the three terminal *n*FET synapse's drain-to-channel depletion region is one-sided: 95% of the space-charge layer appears on

the drain side of the junction. When V_{dc}=30V, peak field occurs 0.14 μm into this space-charge layer. A hot-electron population therefore is available near the channel edge of the space-charge layer. By contrast, in the conventionally doped well–drain transistor, the drain-to-channel depletion region is symmetric rather than one-sided; peak field does not occur until 2μm into the space-charge layer.

Second, the higher surface-acceptor concentration raises the synapse transistor's threshold voltage V_t from about 0.8V to about 6V. It is evident from Figure 2.12 that electron transport within the SiO_2 depends on the direction of the oxide electric field. Where the floating-gate voltage exceeds the surface potential, the oxide electric field sweeps injected electrons across the SiO_2 to the floating gate; where the surface potential exceeds the floating-gate voltage, injected electrons tend to return to the silicon surface. When V_{dc}=30V, the synapse transistor's conduction-band potential is 3.2V at z=0.22μm, whereas the surface potential does not exceed the floating-gate voltage until z=0.37μm. The gate current arises primarily in the intervening region (0.22<z<0.37μm). By contrast, in the conventional well–drain transistor with V_{dc}=30V, the conduction-band potential does not reach 3.2V until 0.9μm into the space-charge layer. Here the surface potential exceeds the gate voltage by 6.5V, preventing a gate current.

To measure the CHEI, I fabricated the synapse of Figure 2.11 without a tunneling junction. I show a CHEI efficiency plot in Figure 2.13. I plot the data as efficiency because the gate current increases linearly with the source current over the subthreshold and perithreshold source-current ranges (see Figure 2.15), where I define perithreshold as that regime where the dependency of the MOS channel current on the floating-gate voltage transitions from exponential to square-law behavior. In Figure 2.14, I empirically fit the measured CHEI efficiency versus (part A) the drain-to-channel potential and (part B) the gate-to-channel potential; my colleague Paul Hasler is investigating the relevant high-field transport physics to derive equivalent analytic results [12]. As I did for the four-terminal *n*FET synapse, I reference the drain to the channel potential because the hot-electron population derives from the drain-to-channel electric field, and I reference the floating gate to the channel potential because the direction of electron transport within the oxide derives from the direction of the gate-to-channel electric field.

In the three-terminal *n*FET synapse, the drain-to-channel electric field increases with the drain voltage; consequently, the gate current also increases with the drain voltage. Part A of Figure 2.14 shows the CHEI efficiency versus the drain-to-channel potential. These data are fit by

$$I_g = \beta I_s e^{-\left(\frac{V_\beta}{V_{dc}+V_\eta}\right)^2} \tag{2.22}$$

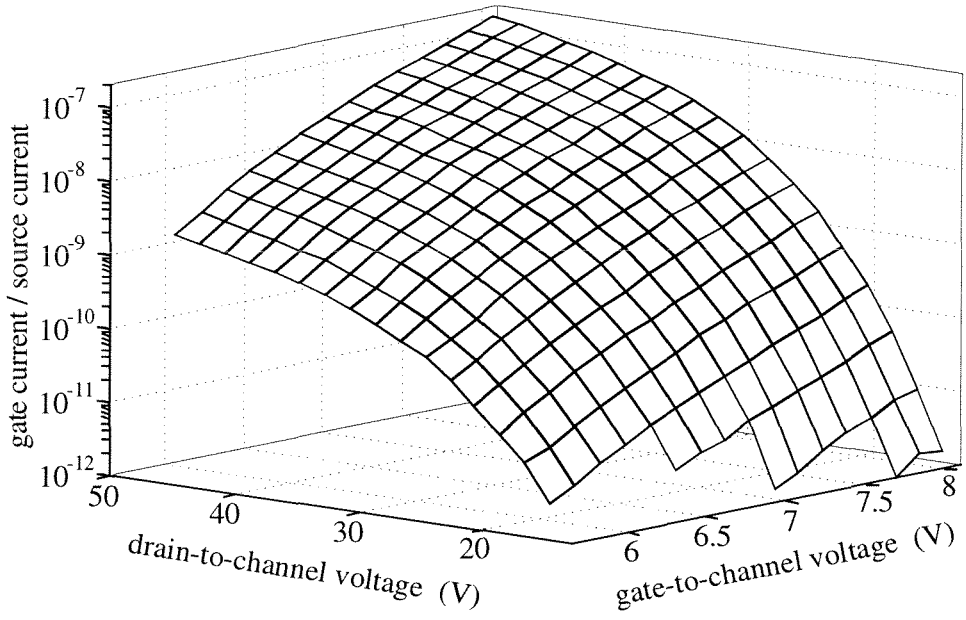


Figure 2.13 Three-terminal *n*FET-synapse CHEI surface plot. I measured the CHEI gate current I_g versus both the gate-to-channel potential V_{gc} and the drain-to-channel potential V_{dc}, for a fixed source current I_s=2μA. I plotted the gate current I_g divided by the source current I_s. In both the subthreshold and the perithreshold regimes, I_g increases linearly with I_s (see Figure 2.15); consequently, these data show the CHEI efficiency for both regimes. The RMS deviation between these data and Eqn. (2.24) is 1.2×10^{-9}, with η=3.63, and the other fit parameters as shown in Figure 2.14.

where I_g is the gate current; I_s is the source current; V_{dc} is the drain-to-channel potential; and β, V_β, and V_η are fit constants.

In Figure 2.12, I define z' to be that location where the oxide electric field is zero. Because z' increases with the floating-gate voltage, the gate current also increases with the floating-gate voltage. Part B of Figure 2.14 shows the CHEI efficiency versus the gate-to-channel potential. These data are fit by

$$I_g = \alpha I_s e^{-\frac{V_\alpha}{V_{gc}}} \tag{2.23}$$

where V_{gc} is the floating-gate–to–channel potential, and α and V_α are fit constants.

I incorporate the drain-voltage and gate-voltage dependencies of Eqns. (2.22) and (2.23), respectively, into a final CHEI equation:

$$I_g = \eta I_s e^{-\frac{V_\alpha}{V_{gc}} - \left(\frac{V_\beta}{V_{dc}+V_\eta}\right)^2} \tag{2.24}$$

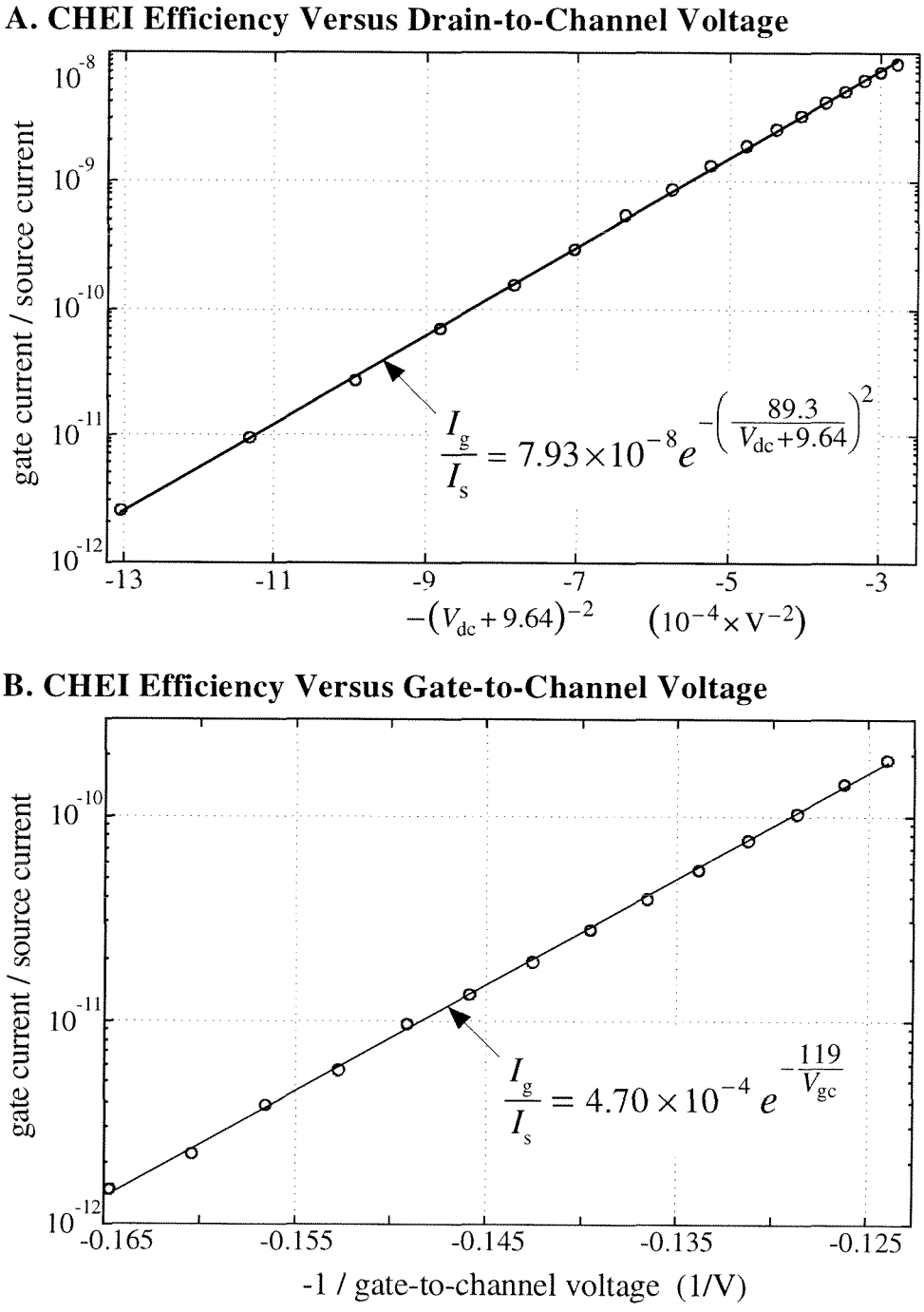


Figure 2.14 Three-terminal *n*FET-synapse CHEI dependencies. (A) I fixed the gate-to-channel voltage at V_{gc}=6.7V and the source current at I_s=2μA, and plotted the CHEI efficiency versus the drain-to-channel voltage V_{dc}. (B) I fixed the drain-to-channel voltage at V_{dc}=20V and the source current at I_s=2μA, and plotted the CHEI efficiency versus the gate-to-channel voltage V_{gc}.

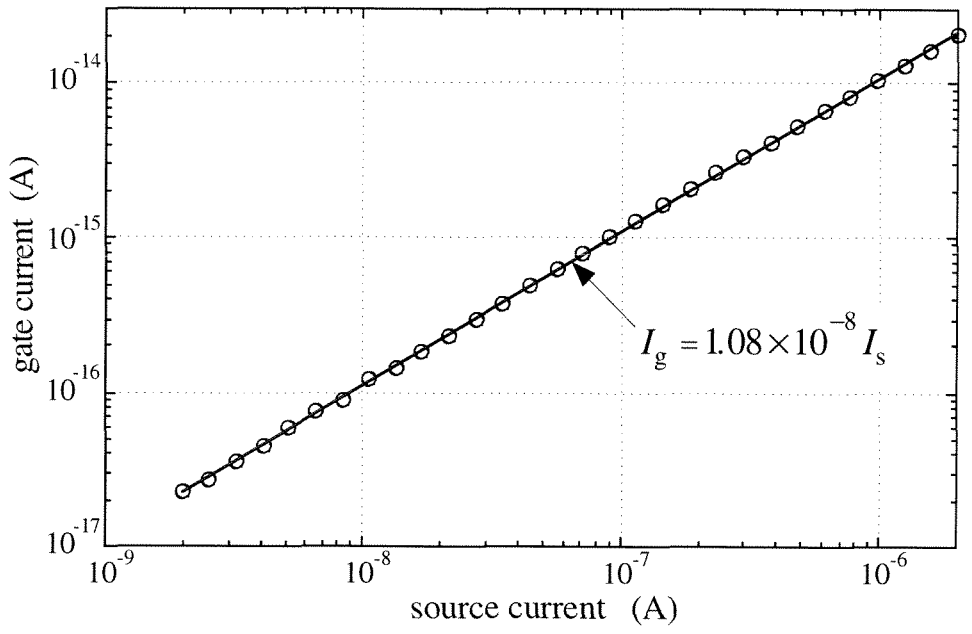


Figure 2.15 Three-terminal *n*FET-synapse gate current versus source current. I held the drain-to-bulk and gate-to-bulk voltages fixed at V_{db}=35 V and V_{gb}=7 V, respectively, and measured the gate current I_g versus the source current I_s. These data show that the three-terminal *n*FET synapse's CHEI efficiency is independent of source current over both the subthreshold and perithreshold source-current ranges.

where η is a fit constant; and V_α, V_β, and V_η remain unchanged from Eqns. (2.22) and (2.23). CHEI in the three-terminal *n*FET synapse is about 10^3 times less efficient than it is in the four-terminal *n*FET synapse (compare Figure 2.3 and Figure 2.13); this difference is a consequence of the three-terminal synapse's lower drain doping (compare part C of Figure 2.1 with part A of Figure 2.12).

2.2.3 The Gate-Current Equation

Because the tunneling and CHEI gate currents flow in opposite directions, I obtain a final gate-current equation by subtracting Eqn. (2.24) from Eqn. (2.1):

$$I_g = I_{to} e^{-\frac{V_f}{V_{ox}}} - \eta I_s e^{-\frac{V_\alpha}{V_{gc}} - \left(\frac{V_\beta}{V_{dc}+V_\eta}\right)^2} \tag{2.25}$$

This equation describes the three-terminal *n*FET synapse's gate current over the subthreshold and perithreshold source-current ranges. This synapse exhibits four operating regions, depending on the drain-to-channel voltage:

1. **V_{dc}<10 V:** The tunneling and CHEI gate currents both are small; the weight W is nonvolatile.

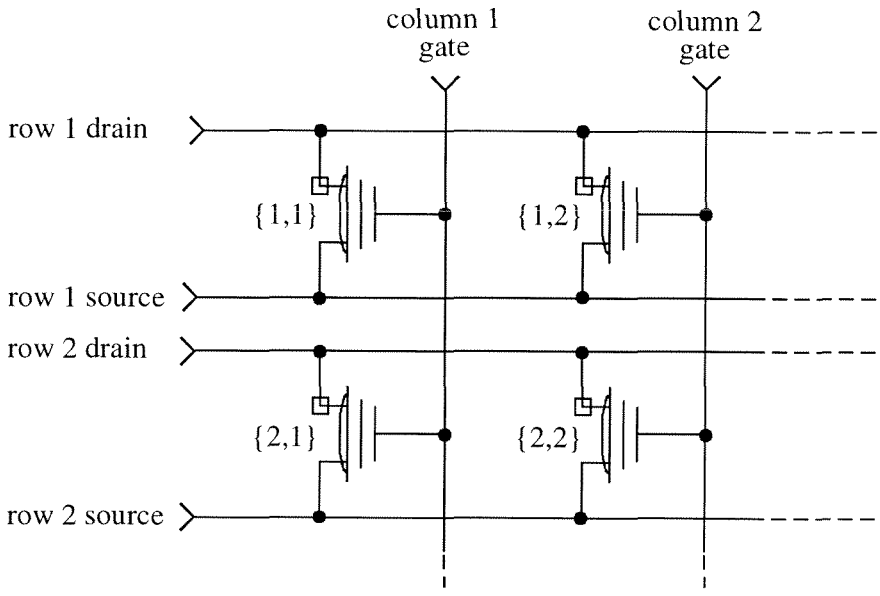


Figure 2.16 A 2×2 array of three-terminal *n*FET synapses. The box on the transistor's drain terminal denotes a well–drain; the curved line in the transistor symbol denotes a pbase channel implant. The row synapses share common drain wires; consequently, tunneling or injection at one row synapse can cause undesired tunneling or injection at other row synapses.

2. **10V<V_{dc}<30V:** The tunneling gate current is small, but the CHEI gate current is not small; electrons are added to the floating gate, decreasing the weight *W*.

3. **30V<V_{dc}<40V:** Neither the tunneling nor the CHEI gate currents are small; the floating-gate asymptotes to a voltage where the gate current of Eqn. (2.25) is zero.

4. **V_{dc}>40V:** The tunneling gate current is larger than the CHEI gate current; electrons are removed from the floating gate, increasing the weight *W*. Although drain voltages that transiently exceed 40V are useful for learning, drain voltages that continuously exceed 40V can lead to excessive power dissipation, damaging the synapse.

2.2.4 Isolation and Weight Updates in a Synaptic Array

As I did for the four-terminal *n*FET synapses, I fabricated a simplified 2×2 array of three-terminal *n*FET synapses to investigate isolation during tunneling and injection, and to measure the synapse weight-update rates. Because this 2×2 array uses the same row–column addressing

Table 2.2 Three-terminal *n*FET-synapse array terminal voltages. I applied these voltages to the array of Figure 2.16, to obtain the data in Figure 2.17.

	column 1 gate	column 2 gate	row 1 drain	row 2 drain	row 1 source	row 2 source
read	+5	0	+5	0	0	0
tunnel	0	+4.5	+35	0	+2	0
inject	+5	0	+25	0	0	0

that I will employ in larger arrays, it allows me to characterize the synapse isolation and weight-update rules completely.

I show the array in Figure 2.16. I chose, from among the many possible ways of using the array, to select source current as the synapse output, and to turn off the synapses during tunneling. I applied the voltages shown in Table 2.2 to read, tunnel, or inject synapse {1,1} selectively, while ideally leaving the other synapses unchanged.

2.2.4.1 Synapse Isolation

A three-terminal *n*FET synapse both tunnels and injects from its drain terminal; because the drains of the array-synapse transistors connect within rows, but not within columns, the tunneling and injection crosstalk between column synapses is negligible. The crosstalk between row synapses should decrease exponentially with the voltage differential between the floating gates of the selected and deselected synapses, just like in the four-terminal *n*FET array. Unfortunately, when the three-terminal *n*FET synapse's drain exceeds about 35V, self-limiting avalanche (*pn*) breakdown occurs at the n^+ tunneling implant. I describe this breakdown process in Section 2.2.6. If the floating-gate voltage is high, then hot electrons generated by this avalanche-breakdown process can inject onto the floating gate, causing a gate current not included in Eqn. (2.24).

When I inject a row synapse, I use 5V control-gate inputs; the voltage differential between the floating gates of the selected and deselected synapses is about 4V, and the crosstalk between the row synapses is <0.01%. When I tunnel a row synapse, however, I can cause both tunneling and avalanche injection at the other row synapse. If I tunnel the {1,1} synapse, and the floating-gate voltage of the {1,2} synapse is high, then electrons inject onto the {1,2} synapse's floating gate by means of avalanche injection. If, on the other hand, the floating-gate voltage of the {1,2} synapse is low, then electrons tunnel off the {1,2} synapse's floating gate.

I show three-terminal *n*FET-synapse-isolation data in Figure 2.17. To obtain the data in part A, I first initialized all four synapses to I_s=30nA. I then tunneled the {1,1} synapse up to 2μA, and injected it back down to 30nA, while I measured the source currents of the other three synapses. As I expected, the row 2 synapses were unaffected by either the tunneling or the injection. The {1,2} synapse was similarly unaffected by the injection, but during tunneling experienced both avalanche injection and parasitic tunneling. A 4.7V signal on the column 2 gate input exactly balanced these parasitic effects; unfortunately, this optimum gate voltage varied with the {1,2} synapse's weight value. I used a 4.5V control-gate input, so parasitic tunneling slightly exceeded avalanche injection at the {1,2} synapse.

To obtain the data in part B of Figure 2.17, I first initialized all four synapses to I_s=2μA. I injected the {1,1} synapse down to 30nA, and then tunneled it back up to 2μA. As in the experiment of part A, when the {1,1} synapse tunneled, the {1,2} synapse experienced both avalanche injection and parasitic tunneling. A 4.3V control-gate input exactly balanced these parasitic effects. With my chosen 4.5V control-gate signal, avalanche injection slightly exceeded parasitic tunneling at the {1,2} synapse.

The measured crosstalk between the row synapses was ~0.5% during tunneling, and <0.01% during CHEI. As I describe in Section 2.3, I anticipate that I can achieve <0.01% crosstalk for both operations when I fabricate my synapses in a more modern process.

2.2.4.2 Synapse Weight Updates

The source currents that I used for the synapse-isolation measurements (see Figure 2.17) extended into the perithreshold regime. I chose perithreshold source currents to speed the CHEI process, because CHEI in the three-terminal *n*FET synapse is about 10^3 times less efficient than it is in the four-terminal *n*FET synapse. This change affects the weight-update rule significantly, as I now describe.

I repeated the experiment of Figure 2.17 (part A), for several tunneling and injection drain voltages; in Figure 2.18, I plot the magnitude of the temporal derivative of the source current versus the source current, for a synapse transistor with (part A) a set of fixed tunneling voltages, and with (part B) a set of fixed drain voltages. I now derive a weight-update rule to fit these data.

2.2.4.2.1 The Tunneling Weight-Increment Rule

Tunneling in the three-terminal *n*FET synapse is functionally identical to tunneling in the four-terminal *n*FET synapse; consequently, the subthreshold weight-increment rule is the same:

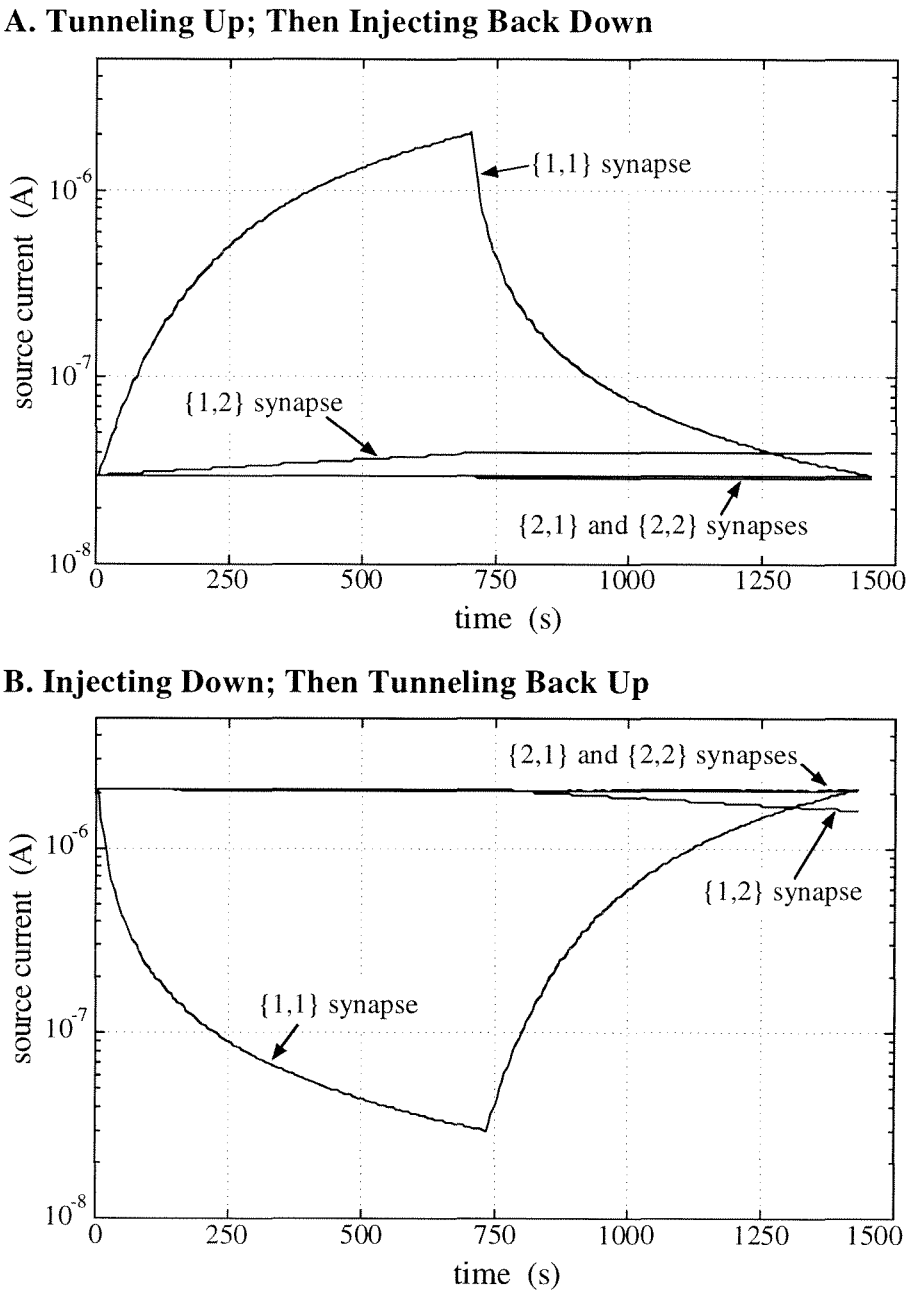


Figure 2.17 Isolation in a 2×2 array of three-terminal *n*FET synapses. The terminal voltages for both experiments are shown in Table 2.2 (see pg. 40). (A) I first tunneled the {1,1} synapse up to 2μA, then I injected it back down to 30nA, while I measured the source currents of the other three synapses. Crosstalk to the {1,2} synapse, defined as the fractional change in the {1.2} synapse's source current divided by the fractional change in the {1,1} synapse's source current, was 0.52% during tunneling, and was 0.023% during injection. (B) I first injected the {1,1} synapse down to 30nA, then I tunneled it back up to 2μA. Crosstalk to the {1,2} synapse was 0.001% during injection, and was 0.43% during tunneling.

$$\frac{\partial W}{\partial t} \approx \frac{1}{\tau_{tun}} W^{(1-\sigma)} \tag{2.7}$$

where the parameters σ and τ_{tun} remain as defined in Eqns. (2.8) and (2.9), respectively. During tunneling, a three-terminal *n*FET synapse also experiences an avalanche-induced CHEI gate current that is not present in the four-terminal device. In most cases, this gate current is small when compared with the tunneling gate current, and can be ignored safely.

For subthreshold source currents, Eqn. (2.7) models the three-terminal *n*FET-synapse weight-increment data accurately; for perithreshold source currents, however, the fit is poor. In perithreshold, a synapse's source current no longer increases exponentially with the floating-gate charge Q_{fg}; rather, the source current increases more slowly, as the MOSFET transitions to a square-law device. Consequently, the synapse's behavior no longer follows Eqn. (1.2). Although the synapse's output still is a graded, analog signal, the synapse no longer performs a multiply. Although this may limit potential applications, in practice I expect that most learning systems will use binary-valued (action-potential) control-gate inputs; for these applications, the graded output is of primary importance, whereas the multiply is inconsequential.

Rather than redefining W for perithreshold operation, I instead assume that W still increases exponentially with Q_{fg} (recall that $W \equiv \exp(Q_{fg}/Q_T)$). Consequently, the synapse's weight-increment rate decreases for perithreshold source currents, not because the tunneling process has changed, but rather because the synapse's source current increases more slowly with Q_{fg}. This effect is clearly evident in part A of Figure 2.18. To model the effect, I extend Eqn. (2.7) empirically, with the following approximation:

$$\frac{\partial W}{\partial t} \approx \frac{1}{\tau_{tun\_3n}} \frac{\Delta W_{max} \times W^{(1-\sigma)}}{W_{corner} + W^{(1-\sigma)}} \tag{2.26}$$

I find the maximum weight change ΔW_{max}, the saturation weight value W_{corner}, and the time constant τ_{tun_3n} by empirical measurement; these parameters all vary with the tunneling voltage V_{tun}. Equation (2.26) fits the three-terminal *n*FET-synapse tunneling weight-increment data accurately, for both subthreshold and perithreshold source currents.

The weight-increment rate saturation that I observed in the three-terminal *n*FET synapse also occurs in the four-terminal *n*FET synapse, when I operate the four-terminal synapse with perithreshold source currents. This observation is consistent with the previous discussion, because both the tunneling process and the weight equation (Eqn. (1.2)) are functionally identical in the three-terminal and four-terminal *n*FET synapses. Interestingly, repeated induction of LTP in neural synapses also causes saturation: The synaptic strength rises to a maximum level, beyond which it cannot be increased experimentally [13].

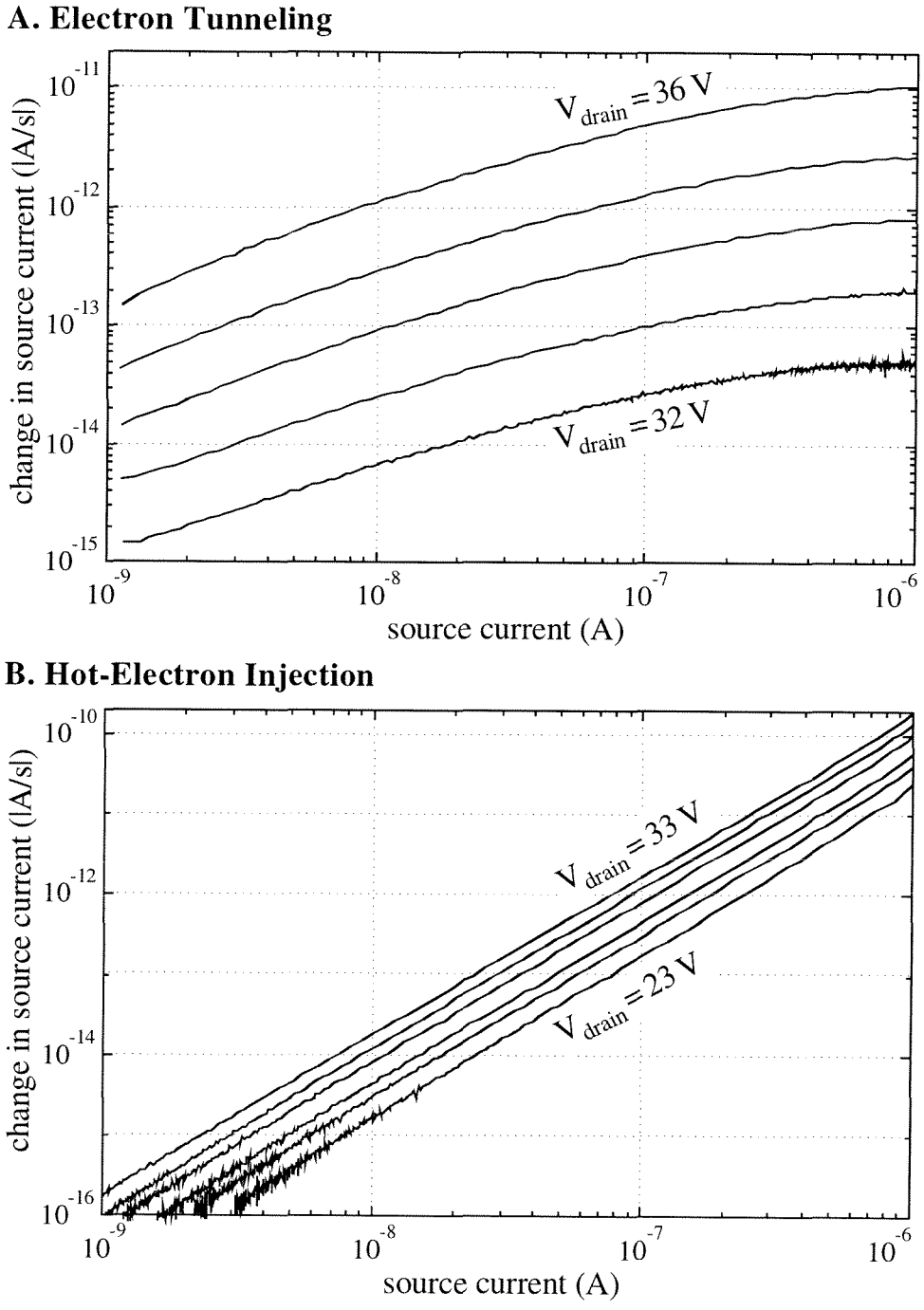


Figure 2.18 Three-terminal *n*FET-synapse (A) tunneling and (B) CHEI weight updates. In both experiments, I measured the synapse's source current I_s versus time, and plotted $|\partial I_s/\partial t|$ versus I_s. I fixed the synapse's terminal voltages; consequently, the change in I_s is a result of changes in the synapse's weight W. In part A, I applied V_{in}=0V, V_s=0V, and stepped V_{ds} from 32V to 36V in 1V increments; in part B, I applied V_{in}=5V, V_s=0V, and stepped V_{ds} from 23V to 33V in 2V increments. I turned off the tunneling and CHEI at regular intervals, to measure I_s. The tunneling weight updates are described by Eqn. (2.26); the CHEI weight updates are described by Eqn. (2.29).

2.2.4.2.2 The CHEI Weight-Decrement Rule

CHEI in the four-terminal *n*FET synapse is different from CHEI in the three-terminal *n*FET synapse, for the three reasons. First, the dependency of the gate current on the floating-gate–to–channel voltage is small in the four-terminal synapse, but is not small in the three-terminal synapse (compare Figure 2.3, at small drain-to-channel voltages, with Figure 2.13). Second, I operate the four-terminal synapse with subthreshold source currents, whereas I operate the three-terminal synapse with both subthreshold and perithreshold source currents. Third, in the four-terminal synapse I can approximate the dependency of the gate current on the drain-to-channel voltage by a simple exponential; in the three-terminal synapse, the drain-voltage range is much larger, and the dependency of the gate current on the drain-to-channel voltage is more complicated (compare Eqns. (2.2) and (2.22)).

I have been unable to derive a closed-form CHEI weight-decrement expression for the three-terminal *n*FET synapse. Although the derivation generally follows that for the four-terminal *n*FET synapse (see Section 2.1.4.2.2), the analysis, unfortunately, does not yield a simple result. However, I can describe the result qualitatively. I begin with the three-terminal *n*FET synapse's CHEI gate-current equation, Eqn. (2.24):

$$I_g = \eta I_s e^{-\frac{V_\alpha}{V_{gc}} - \left(\frac{V_\beta}{V_{dc}+V_\eta}\right)^2} = f\left(V_{gc}, V_{dc}\right) I_s \tag{2.27}$$

I also use the CHEI weight-decrement expression from the four-terminal *n*FET synapse:

$$\frac{\partial W}{\partial t} = -\frac{1}{\tau_{inj}} W^{(2-\varepsilon)} \tag{2.28}$$

For a three-terminal *n*FET synapse with fixed terminal voltages, V_{gc} increases as W increases, whereas V_{dc} decreases as W increases; f, which depends on both, will increase as W increases. If I consider this effect alone, then I expect the form of the CHEI weight-decrement rule in Eqn. (2.17) to change from $W^{(2-\varepsilon)}$ to $W^{(2+x)}$, where x is a positive-valued correction term. However, because I operate the three-terminal *n*FET synapse with perithreshold source currents, the weight-decrement rate decreases with W, not because the CHEI process has changed (any small potential drop along the channel is inconsequential when compared with the large drain-to-channel and gate-to-channel potentials), but rather because the source current decreases more slowly with Q_{fg} than it does in the subthreshold regime. This second effect causes x to decrease. Although I cannot predict a value for x analytically, the data of Figure 2.18 show empirically that the two effects that I have described cancel almost exactly, yielding $x \approx 0$. Consequently, the CHEI weight-decrement rule for the three-terminal *n*FET synapse is given approximately by

$$\frac{\partial W}{\partial t} \approx -\frac{1}{\tau_{inj\_3n}} W^2 \tag{2.29}$$

where I find τ_{inj_3n} by empirical measurement.

2.2.4.2.3 The Synapse Weight-Update Rule

I obtain the complete weight-update rule, for the three-terminal *n*FET synapse, by adding Eqns. (2.26) and (2.29):

$$\frac{\partial W}{\partial t} \approx \frac{1}{\tau_{tun\_3n}} \frac{\Delta W_{max} \times W^{(1-\sigma)}}{W_{corner} + W^{(1-\sigma)}} - \frac{1}{\tau_{inj\_3n}} W^2 \tag{2.30}$$

2.2.5 Impact Ionization

As it does in the four-terminal *n*FET synapse, in the three-terminal *n*FET synapse the drain-to-channel electric field that generates electrons for oxide injection also liberates additional electron–hole pairs, causing I_d to increase exponentially with V_{dc}. I show impact-ionization data from the three-terminal *n*FET synapse in Figure 2.19. Because the drain-voltage range is large, I cannot fit these data accurately using a simple exponential; I instead use a lucky-electron model [9, 10]. To improve the fit, I modify the conventional lucky-electron formulation by subtracting an offset potential, V_γ, from the drain-to-channel voltage:

$$I_d = I_s \left(1 + \gamma e^{\sqrt{\frac{V_m}{V_{dc} - V_\gamma}}} \right) \tag{2.31}$$

where I_d is the drain current; and γ, V_m, and V_γ are empirical constants. Eqn. (2.31) is simpler than, but still is generally consistent with, my colleague Paul Hasler's recent derivation of an improved model for the impact-ionization process [12].

2.2.6 Drain Leakage Current and Avalanche Injection

In the three-terminal *n*FET synapse (see Figure 2.11), a 2μm wide floating-gate extension traverses from the channel–drain edge, over the n^- well–drain, to the n^+ tunneling implant (the n^+ well–drain contact). I placed a field-oxide channel stop beneath this gate extension, ostensibly to prevent the channel–surface depletion layer from reaching the n^+. Unfortunately, in the 2μm Orbit process that I use, the threshold voltage for a field-oxide transistor is $V_t \approx 18$V. In the three-terminal synapse, when $V_{dg} > 15$V, a parasitic *p*-type MOS channel forms in the n^- well, beneath the channel stop. When $V_{dg} > 30$V, *pn* breakdown occurs where the induced *p*-type channel meets the n^+ well contact [2]. I show three-terminal *n*FET-synapse drain-leakage data in Figure 2.20.

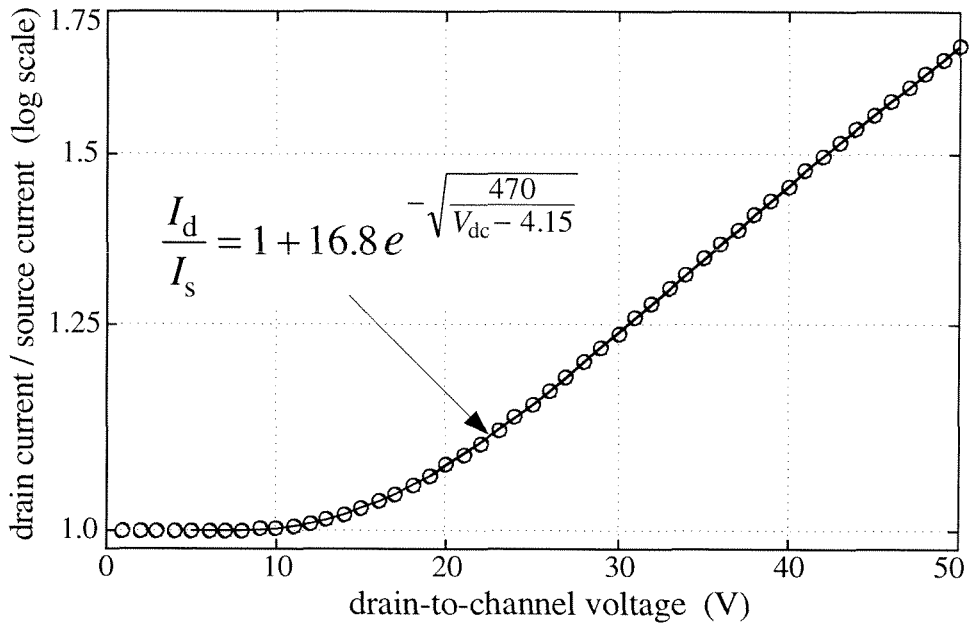


Figure 2.19 Three-terminal *n*FET synapse impact ionization versus drain-to-channel voltage. I fixed the gate-to-channel voltage at $V_{gc}=6.15\,V$ and the source current at $I_s=100\,nA$, and I measured the drain current I_d versus the drain-to-channel voltage V_{dc}. Because this synapse both tunnels and injects from the drain terminal, the drain-voltage range that I anticipate using in my learning systems is large. Consequently, I fit these data over the entire drain-voltage range using a modified lucky-electron model, rather than over a subset of the range using an exponential fit (compare with Figure 2.9).

For drain voltages greater than about 35 V, *pn* breakdown at the n^+ well contact generates free carriers, and the MOS channel beneath the field-oxide channel stop provides a leakage path from the drain contact to the substrate. The field-oxide channel conductance restricts this leakage current; consequently, the breakdown process is self-limiting. Unfortunately, junction breakdown generates hot electrons, thereby inducing a poorly controlled, parasitic hot-electron gate current.

The tunneling implant in the four-terminal *n*FET synapse experiences a leakage phenomenon similar to that observed in the three-terminal *n*FET synapse: When V_{tun} exceeds about 35 V, a leakage current flows from the n^+ tunneling implant to the substrate. I show tunneling-junction leakage data in Figure 2.21. Interestingly, in the four-terminal *n*FET synapse, the avalanche-breakdown process does not induce a parasitic hot-electron gate current, suggesting that, in the three-terminal *n*FET synapse, the avalanche injection occurs near the channel–drain edge, rather than at the tunneling junction. I do not yet understand either the junction breakdown or the parasitic injection completely; I hope simply to eliminate them in future synapses (see Section 2.3.2).

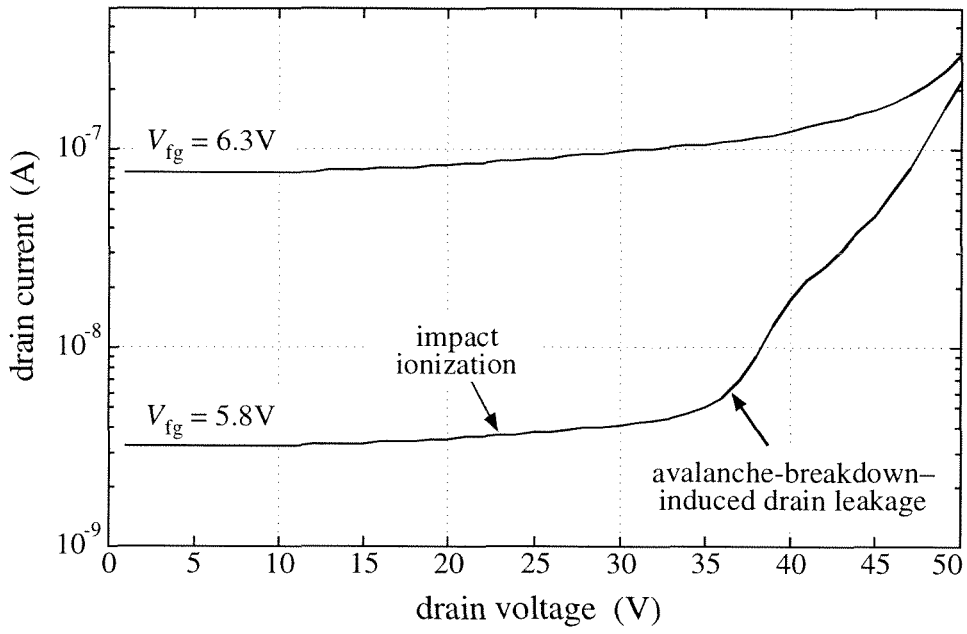


Figure 2.20 Three-terminal *n*FET-synapse drain current versus drain voltage. I held the source at ground, the floating gate at the voltages indicated, and measured the drain current I_d as I swept the drain from 0V to 50V. For V_{dg}>15V, a parasitic *p*-type MOS channel forms in the n^- well–drain, beneath the channel stop. For V_{dg}>30V, *pn* breakdown occurs where the induced *p*-type channel meets the n^+ well contact. Junction breakdown generates free carriers, causing a leakage current to flow from the n^+ well contact to the *p*-type substrate. The field-oxide channel conductance restricts this leakage current; consequently, the breakdown process is self-limiting. Although all *n*-well tunneling junctions fabricated in the 2μm Orbit process exhibit this leakage current (see Figure 2.21), the effect is less important for the four-terminal *n*FET synapse than it is for the three-terminal *n*FET synapse, because the leakage path in the four-terminal synapse is from the tunneling junction to ground, whereas the leakage path in the three-terminal synapse is from the drain to ground. In addition, in the three-terminal synapse, the avalanche-breakdown process causes undesired electron injection onto the floating gate (see Sections 2.2.4.1 and 2.2.6).

2.3 Further Development

My *n*FET synapses already possess those attributes that I believe are essential for building a silicon learning system. They allow nonvolatile analog weight storage, permit simultaneous memory reading and writing, and allow bidirectional weight updates that are a function of both the applied terminal voltages and the present synaptic output. However, further development will improve the devices substantially. I discuss four areas for improvement; in all cases, more modern processing will readily allow these improvements.

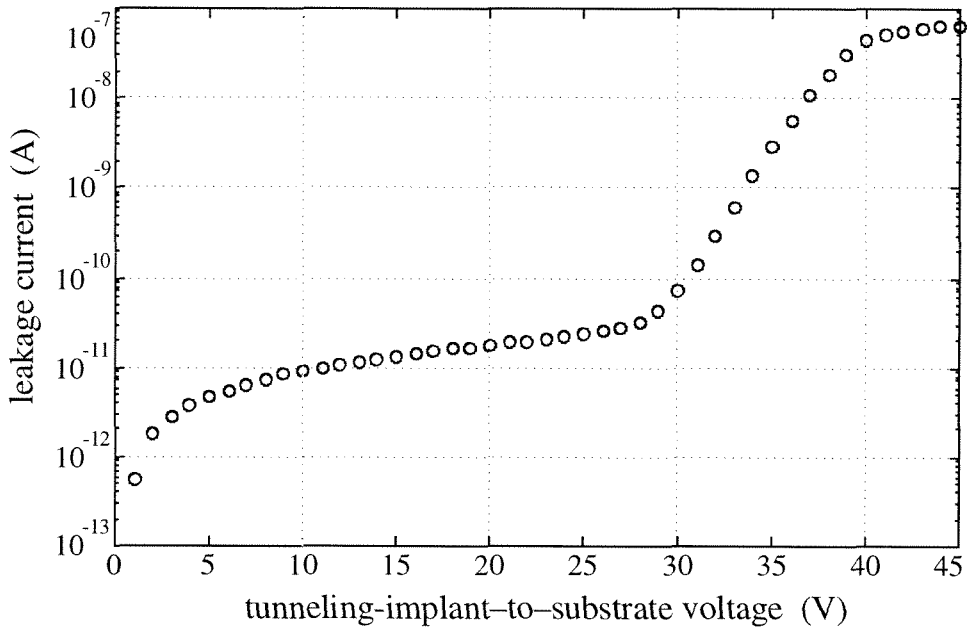


Figure 2.21 Tunneling-implant leakage current versus implant-to-substrate voltage. I held the substrate at ground and the floating gate at 5V, and measured the leakage current from the n^+ tunneling implant to the substrate, as I swept the tunneling implant from 0V to 45V. For $V_{tun}>15$V, a parasitic p-type MOS channel forms in the n^- well, beneath the channel stop. For $V_{tun}>30$V, pn breakdown occurs where the induced p-type channel meets the n^+ well contact. Junction breakdown generates free carriers, causing a leakage current to flow from the n^+ tunneling implant to the p-type substrate. These data suggest strongly that the field-oxide transistor's channel conductance restricts this leakage current. The n^- well is a backgate for the field-oxide transistor; as I increase the well voltage, the field-oxide transistor's channel conductance increases, and the leakage current increases. The subthreshold leakage-current range extends to about 50nA. Above this 50nA threshold, the field-oxide transistor's channel current no longer increases exponentially with the tunneling-implant voltage V_{tun}.

2.3.1 Reduced Tunneling Voltages

All my synapse transistors (nFET and pFET) require large tunneling voltages, because the gate-oxide thickness in the 2μm Orbit process ranges from 350–450Å. More modern processes have thinner oxides and therefore lower tunneling voltages; if I fabricate my synapses in a modern process with 100Å gate oxides, they will tunnel at 12V, instead of at the 35V that I presently require.

2.3.2 Reduced Tunneling-Junction Leakage

I believe that I can eliminate the tunneling-junction leakage, in both the three-terminal and four-terminal synapses, either by using a MOS process with a thicker channel stop, or by using a more modern process with lower tunneling voltages. I expect the three-terminal array synapses to achieve <0.01% tunneling crosstalk, which the four-terminal array synapses already achieve, once I eliminate the leakage pathway and, with it, the avalanche-induced gate current.

2.3.3 Reduced Overlap Capacitances

In a four-terminal *n*FET synapse, the overlap capacitance between the n^- tunneling well and the floating gate is about 5fF. Because the tunneling implant undergoes large voltage swings, the coupling from the n^- well to the floating gate alters the synapse's channel current significantly. Similarly, in the three-terminal *n*FET synapse, the overlap capacitance between the drain and the floating gate is about 5fF. Because V_d ranges from ground to about 45V, drain-to-gate coupling alters this synapse's channel current significantly. In both devices, I minimize the effect by using oversize (1pF) gate capacitors. In future synapses with smaller tunneling junctions (see Section 2.3.4), these overlap capacitances will decrease, and the sensitivity of a synapse's channel current to the tunneling voltage will likewise decrease.

2.3.4 Smaller Synapse Size

Both my four-terminal and my three-terminal *n*FET synapses are large, for two reasons: (1) the n^- tunneling wells are large, and (2) I employ oversize gate capacitors. If I fabricate future synapses in a modern EEPROM process, two improvements are possible. First, EEPROM processes often employ graded [14] tunneling junctions; if I replace my n^- wells with graded junctions, the synapses get smaller, and the overlap capacitances decrease. Second, EEPROM processes have much thinner gate oxides, and, consequently, much lower tunneling voltages. With lower tunneling voltages, the required voltage differential between the floating gates of selected and deselected array synapses can be smaller. Reduced overlap capacitances and reduced floating-gate voltage swings together will allow me to use smaller gate capacitors. For both synapses, these improvements will reduce the synapse size, and speed the weight-update rates. In addition, for the three-terminal synapse, these changes will improve the array isolation.

References

1 M. Lenzlinger and E. H. Snow, "Fowler–Nordheim tunneling into thermally grown SiO_2," *J. of Appl. Phys.*, vol. 40, no. 6, pp. 278–283, 1969.

2 A. S. Grove, *Physics and Technology of Semiconductor Devices,* New York: John Wiley & Sons, 1967.

3 S. M. Sze, *Physics of Semiconductor Devices,* New York: John Wiley & Sons, 1981.

4 C. Mead, "Scaling of MOS technology to submicrometer feature sizes," *J. of VLSI Signal Processing,* vol. 8, no. 6, pp. 9–25, 1994.

5 J. J. Sanchez and T. A. DeMassa, "Review of carrier injection in the silicon/silicon-dioxide system," *IEE Proceedings-G,* vol. 138, no. 3, pp. 377–389, 1991.

6 C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing,* vol. 8, no. 1, pp. 83–114, 1995.

7 A. G. Andreou and K. A. Boahen, "Neural information processing II," in M. Ismail and T. Fiez, eds., *Analog VLSI Signal and Information Processing,* New York: McGraw-Hill, pp. 358–413, 1994.

8 C. Mead, *Analog VLSI and Neural Systems,* Reading, MA: Addison-Wesley, 1989.

9 W. Shockley, "Problems related to *pn* junctions in silicon," *Solid-State Electronics,* vol. 2, no. 1, pp. 35–67, Pergamon Press, 1961.

10 S. Tam, P. Ko, and C. Hu, "Lucky-electron model of channel hot-electron injection in MOSFET's," *IEEE Trans. Electron Devices,* vol. ED-31, no. 9, pp. 1116–1125, Sep. 1984.

11 S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka, "Reliability issues of flash memory cells," *Proc. of the IEEE,* vol. 81, no. 5, pp. 776–787, 1993.

12 P. Hasler, *Foundations of Learning in Analog VLSI,* Ph.D. Thesis, Department of Computation and Neural Systems, California Institute of Technology, Pasadena, CA, 1997.

13 K. Jeffery and I. Reid, "Modifiable neuronal connections: An overview for psychiatrists," *Am. J. of Psychiatry,* vol. 154, no. 2, pp. 156–164, 1997.

14 E. Takeda, C. Yang, and A. Miura-Hamada, *Hot-Carrier Effects in MOS Devices,* San Diego, CA: Academic Press, 1995.

Chapter 3

The *p*-Type Synapse Transistors

In this chapter, I describe the layout, characteristics, and weight-update behavior of my *p*FET synapse transistors. I describe first a four-terminal *p*FET synapse; from this device, I then develop a more compact guarded-*p*FET synapse.

3.1 A Four-Terminal *p*FET Synapse

The four-terminal *p*FET synapse, shown in Figure 3.1, is a *p*-type MOSFET with a poly1 floating gate, a poly2 control gate, and a fourth terminal used for gate-oxide tunneling. Because the 2μm process that I use is *n*-well, I fabricate the transistor within an *n*-type well in the *p*-type substrate. In the four-terminal *p*FET synapse, the transistor's *n*-well and the tunneling *n*-well are separate. In the guarded-*p*FET synapse (see Section 3.2), I combine these wells to obtain a more compact layout. I use FN tunneling to remove electrons from the floating gate, and use hot-electron injection to add electrons to the floating gate. The four-terminal *p*FET synapse has the following features:

- Electrons tunnel from the floating gate to the tunneling implant through the 400 Å gate oxide. The tunneling implant is identical to that used in the four-terminal *n*FET synapse (see Section 2.1.1). As in the *n*FET synapse, tunneling removes electrons from the floating gate. However, because the *p*FET and *n*FET synapses are complementary, tunneling has the opposite effect on the *p*FET synapse: It decreases, rather than increases, the synapse weight *W*.

- Electrons inject from the drain-to-channel space-charge region to the floating gate. I generate the electrons for oxide injection by means of hole impact ionization in the drain-to-channel depletion region of a subthreshold MOSFET. Channel holes, accelerated in the drain-to-channel electric field, collide with the semiconductor lattice to produce additional electron–

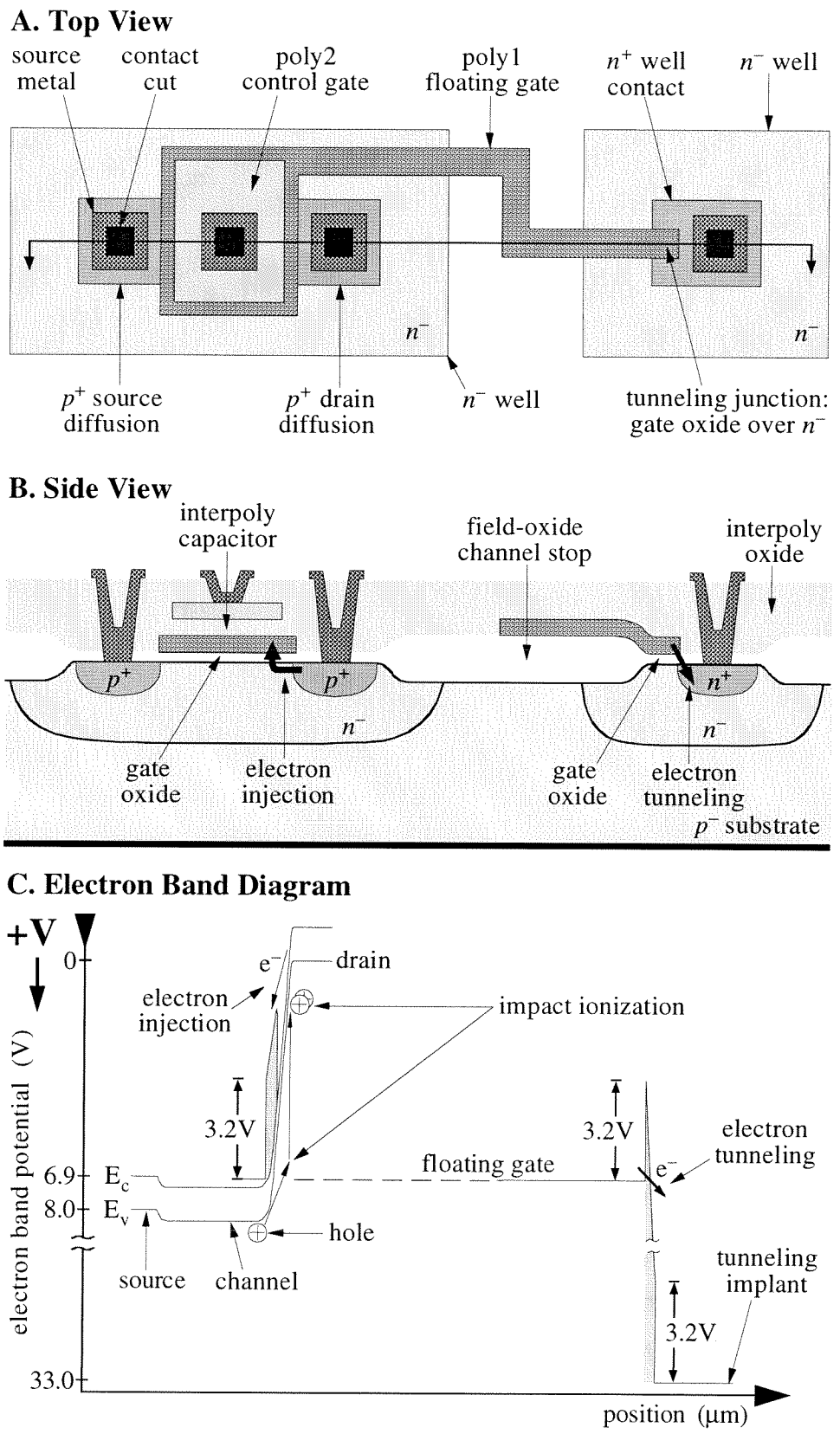


Figure 3.1 The four-terminal *p*FET synapse, showing the electron tunneling and injection locations. The

well contact is not shown. As I did in Figure 2.1, I have aligned the three diagrams vertically, drawn diagrams A and C to scale, exaggerated the vertical scale in diagram B, referenced the voltages in the band diagram to the source potential, drawn the gate-oxide band diagram in the channel direction, and assumed subthreshold source currents (I_s<100nA). In the 2μm Orbit process, the synapse length is 56μm, and the width is 16μm. With a 50fF interpoly capacitor as shown, the coupling coefficient between the poly2 control gate and the poly1 floating gate is only 0.25. I enlarged the interpoly capacitor to 1pF in the test device, thereby increasing the coupling to 0.8. Whereas the tunneling process is identical to that used in the *n*FET synapses (see Section 2.1.1), the injection process is different. I generate electrons for oxide injection by means of hole-impact ionization in the transistor's drain-to-channel depletion region (see Section 3.1.2 for a description of the injection process).

hole pairs. The liberated electrons, promoted to their conduction band by the collision, are expelled rapidly from the drain region by this same drain-to-channel electric field. Electrons expelled with more than 3.2eV of kinetic energy can, if scattered upward into the gate oxide, inject onto the floating gate. As in the *n*FET synapse (see Section 2.1.2), injection adds electrons to the floating gate. Because the transistor is a *p*FET, however, injection increases, rather than decreases, the synapse weight *W*.

- Like the *n*FET synapse, the *p*FET synapse uses the thermally grown gate oxide for all SiO_2 carrier transport.

3.1.1 Electron Tunneling Decreases the Weight

I show the tunneling junction in parts A and B of Figure 3.1, and the energy-band diagram for the tunneling process in part C of Figure 3.1. The layout and band diagram are identical to those for the four-terminal *n*FET synapse (compare Figure 3.1 with Figure 2.1). For both the *n*FET and the *p*FET synapses, I apply positive high voltages to the tunneling implant to remove electrons from the floating gate, thereby increasing the floating-gate voltage. In the *n*FET, increasing the floating-gate voltage increases the channel current and the synapse weight *W*. In the *p*FET, tunneling has the opposite effect: It decreases both the channel current and the weight *W*.

3.1.2 Electron Injection Increases the Weight

I increase the synapse weight *W* by injecting electrons onto the floating gate. Because a *p*FET's channel current comprises holes, *p*FET injection is different from *n*FET injection. I show the injection process in the energy-band diagram of Figure 3.1. I accelerate channel holes in the

drain-to-channel electric field of a subthreshold *p*FET. A fraction of these holes collide with the semiconductor lattice at energies sufficient to liberate additional electron–hole pairs. The ionized electrons, promoted to their conduction band by the collision, are expelled from the drain by this same drain-to-channel electric field. If the electrons are expelled with more than 3.2eV of kinetic energy, they can inject over the 3.2V Si–SiO_2 work-function barrier, into the oxide conduction band. These electrons then are swept over to the floating gate by the oxide electric field. I call the process impact-ionized hot-electron injection (IIHEI). My colleagues and I are the first researchers to describe IIHEI as a means for writing a floating-gate MOS memory [1].

As I described in Sections 2.1.2 and 2.2.2, electron injection, in the *n*FET synapses, is successful only when three conditions are satisfied: (1) the electrons must possess the 3.2eV required to surmount the Si–SiO_2 work-function barrier, (2) the electrons must scatter upward into the gate oxide, and (3) the oxide electric field must be oriented in the proper direction to transport the injected electrons to the floating gate. These same conditions apply to a *p*FET synapse.

A *p*FET, like an *n*FET, readily satisfies requirements 1 and 2. I merely operate the transistor in the subthreshold regime, with a sufficient drain-to-channel voltage. In the *p*FET synapse, this minimum drain-to-channel voltage is about 6.5V, to achieve an IIHEI efficiency of 10^{-10}. In the four-terminal *n*FET synapse, this minimum drain-to-channel voltage is about 2.5V, to achieve a CHEI efficiency of 10^{-10}. The *p*FET synapse's higher drain-voltage requirement, when compared with the *n*FET, arises for three reasons. First, for any given drain-to-channel voltage, the *n*FET experiences a higher drain-to-channel electric field than does the *p*FET, as a result of the bulk *p*-type implant that I add to the *n*FET synapse's channel region. Second, the phonon loss mechanisms in silicon are naturally more efficient for holes (the *p*FET charge carriers) than they are for electrons (the *n*FET charge carriers). Consequently, the *p*FET synapse requires a higher drain-to-channel electric field than does the *n*FET, to achieve the same hot-carrier population. Third, the *p*FET synapse's two-step injection process requires more energy.

A subthreshold *p*FET, unlike a subthreshold *n*FET, satisfies requirement 3 without an additional channel implant. In a subthreshold *p*FET, the gate-to-source voltage typically is less than 1V. During IIHEI, the drain-to-source voltage exceeds 6V. Consequently, the floating gate is at least 5V higher than the drain, and the oxide electric field transports the injected electrons to the floating gate. Unlike conventional *n*FET transistors, conventional *p*FET transistors inject electrons onto their floating gates naturally (at sufficient drain-to-source voltages); they do not need special channel implants to facilitate injection.

In Figure 3.2, I show the four-terminal *p*FET synapse's IIHEI efficiency (gate current I_g divided by source current I_s) versus both the drain-to-channel and the gate-to-channel potentials. I

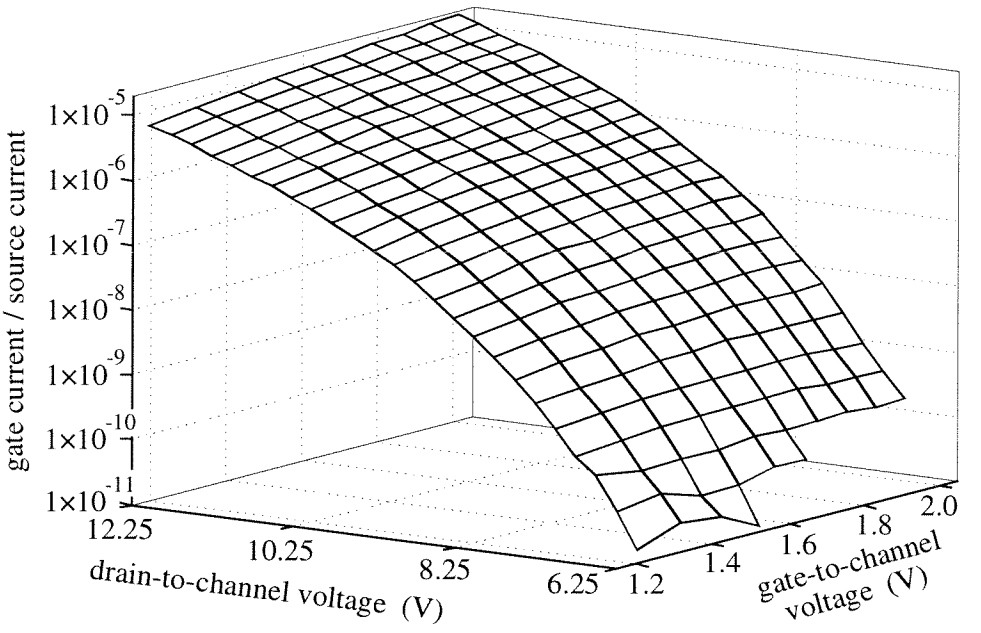


Figure 3.2 Four-terminal *p*FET-synapse IIHEI surface plot. I measured the IIHEI gate current I_g versus both the gate-to-channel potential, V_{gc}, and the drain-to-channel potential, V_{dc}, for a fixed source current I_s=10nA. I plotted the gate current I_g divided by the source current I_s. In the subthreshold regime, I_g increases linearly with I_s (see Figure 3.4); consequently, these data show the IIHEI efficiency for the entire subthreshold source-current range. The IIHEI efficiency increases with both the drain-to-channel and gate-to-channel voltages, for reasons that I discuss in Section 3.1.2. In Figure 3.3, I fit the IIHEI efficiency versus V_{dc}. I anticipate that, for most learning applications, V_{gc} will vary by less than 100mV. Because the IIHEI efficiency depends only weakly on V_{gc}, in fit Eqn. (3.1) I assume that the IIHEI efficiency depends only on V_{dc}.

plot the data as efficiency because the IIHEI gate current increases linearly with the transistor's source current (see Figure 3.4); predictably, because the gate current derives from the impact-ionized electron population, and this electron population increases linearly with the source current. I reference the drain to the channel potential because the hot-electron population derives from the drain-to-channel electric field. I reference the floating gate to the channel potential because the oxide-barrier shape varies with the relative potentials of the channel and the floating gate.

The reference potential for the channel side of the oxide barrier is the electron conduction-band edge; the reference potential for the gate side of the oxide barrier is the floating-gate voltage. As in the three-terminal *n*FET synapse, in the *p*FET synapse electrons inject onto the floating gate over only a restricted range of channel positions z (compare part C of Figure 3.1 with

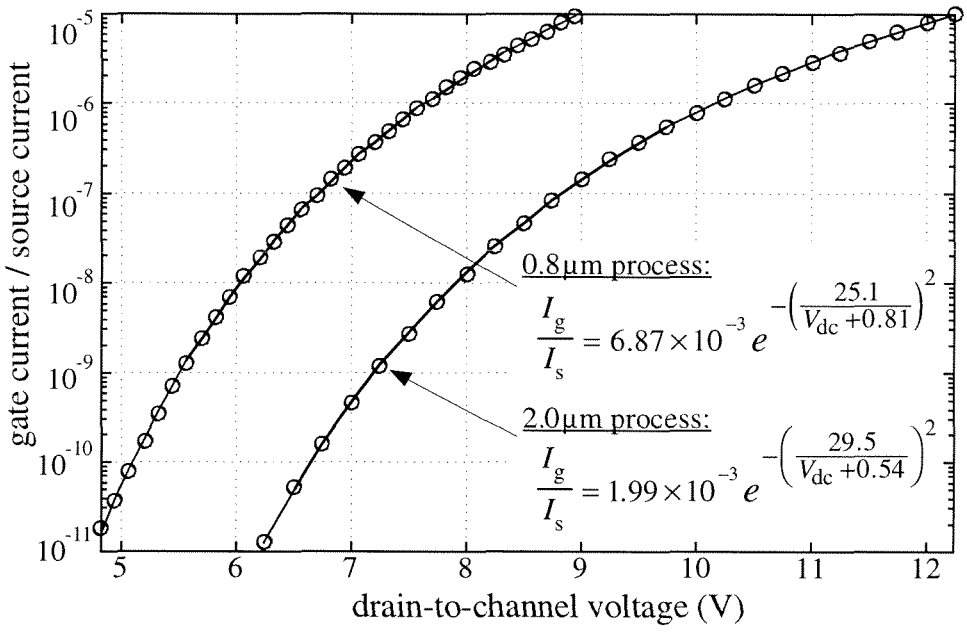


Figure 3.3 Floating-gate *p*FET IIHEI efficiency versus drain-to-channel voltage. I show data from floating-gate pFETs fabricated in 2μm and 0.8μm processes, to illustrate that (1) the IIHEI mechanism is invariant with process linewidth, and (2) the V_{dc} required for IIHEI decreases with process linewidth. During both experiments, I held the gate-to-channel voltages V_{gc} and the source currents I_s fixed, and I measured the gate current I_g versus the drain-to-channel voltage V_{dc}. For the 2μm *p*FET, V_{gc}=1.95V and I_s=10nA; for the 0.8μm *p*FET, V_{gc}=0.9V and I_s=100nA. My empirical fit holds for the entire subthreshold source-current range.

part A of Figure 2.12). Where the floating-gate voltage exceeds the surface potential, the oxide electric field sweeps injected electrons across the SiO_2 to the floating gate; where the surface potential exceeds the floating-gate voltage, injected electrons return to the silicon surface. Fortunately, because the four-terminal *p*FET synapse's floating-gate voltage is high, the dependency of the gate current on the gate-to-channel potential is small. Consequently, I ignore this gate-to-channel dependency in my IIHEI weight-increment rule derivation (see Section 3.1.4.2.2).

In Figure 3.3, I plot the measured IIHEI efficiency, with empirical fits, both for a synapse transistor fabricated in the 2μm process, and for a *p*-type floating-gate MOSFET fabricated in a more modern 0.8μm process. These data show clearly that the results that I derive from the 2μm process scale directly to more modern processes. For the 2μm synapse, when V_{dc} is less than 5V, the IIHEI gate current is exceedingly small, and the weight W remains nonvolatile. When V_{dc} exceeds 6V, the IIHEI gate current causes measurable changes in the synapse weight W. I anticipate using a larger drain-voltage range in my *p*FET-based learning systems than in my four-terminal

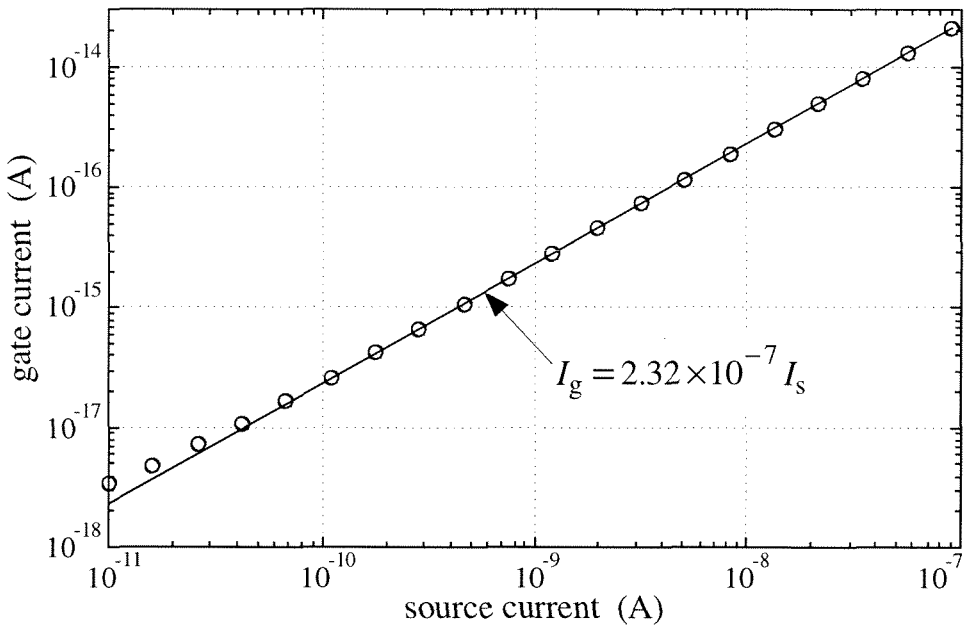


Figure 3.4 Four-terminal *p*FET-synapse gate current versus source current. I held the drain-to-well and gate-to-well voltages fixed at V_{dw}=9.5 V and V_{gw}=1 V, respectively, and measured the gate current I_g versus the source current I_s. For I_s<40 pA, the measured I_g exceeds the fit, as a result of background (thermally generated) electrons that accelerate in the drain-to-channel electric field and inject onto the floating gate. Background injection is not apparent in the data of Figure 2.5 and Figure 2.15, because the minimum gate current in both these figures exceeds 1×10^{-17} A—well above the background injection rate. Background injection occurs in all the synapse transistors, and can be eliminated only by reduction of the drain-to-channel voltage.

*n*FET-based systems, because the *p*FET synapse's gate current increases more slowly with drain-to-channel voltage than does the *n*FET synapse's gate current. Consequently, I fit the *p*FET IIHEI data over the entire drain-voltage range, rather than over only a subset of the range as I did for the *n*FET synapse (compare Figure 3.3 with Figure 2.4).

My empirical fit to the IIHEI data of Figure 3.3 is

$$I_g = \beta I_s e^{-\left(\frac{V_\beta}{V_{dc}+V_\eta}\right)^2} \tag{3.1}$$

where I_g is the gate current; I_s is the source current; V_{dc} is the drain-to-channel potential; and β, V_β, and V_η are fit constants. Both this *p*FET synapse and the three-terminal *n*FET synapse inject electrons from the lightly doped side of their respective drain-to-channel junctions; interestingly, the drain-to-channel voltage dependency of the electron injection process is identical for both devices (compare Eqns. (3.1) and (2.22)).

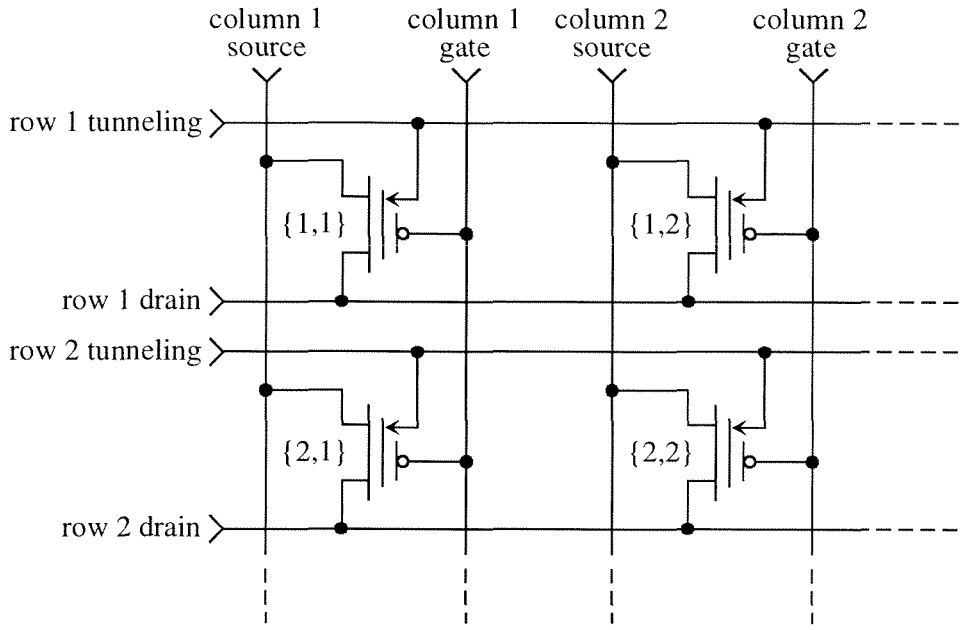


Figure 3.5 A 2×2 array of four-terminal *p*FET synapses. The arrow at each synapse's floating gate denotes a tunneling junction. The well connections are not shown. The row synapses share common tunneling and drain wires; consequently, tunneling or injection at one row synapse can cause undesired tunneling or injection at other row synapses.

3.1.3 The Gate-Current Equation

Because the tunneling and IIHEI gate currents flow in opposite directions, I obtain a final gate-current equation by subtracting Eqn. (3.1) from Eqn. (2.1):

$$I_g = \mathrm{I}_{to} e^{-\frac{\mathrm{V_f}}{V_{ox}}} - \beta I_s e^{-\left(\frac{\mathrm{V_\beta}}{V_{dc}+\mathrm{V_\eta}}\right)^2} \tag{3.2}$$

This equation describes the four-terminal *p*FET synapse's gate current over the entire drain-voltage and subthreshold channel-current ranges.

3.1.4 Isolation and Weight Updates in a Synaptic Array

For this *p*FET synapse, like for the *n*FET devices, a synaptic array can form the basis of a silicon-learning system. I fabricated a simplified 2×2 array of four-terminal *p*FET synapses to investigate isolation during tunneling and injection, and to measure the synapse weight-update rates. I show the array in Figure 3.5. I grounded the *p*-type substrate, applied +12V to the *n*-type well, and referenced all terminal voltages to the well potential. I chose source current as the

Table 3.1 Four-terminal *p*FET-synapse array terminal voltages. I applied these voltages to the array of Figure 3.5, to obtain the data in Figure 3.6.

	column 1 gate	column 2 gate	column 1 source	column 2 source	row 1 drain	row 2 drain	row 1 tunnel	row 2 tunnel
read	–5	0	0	0	–5	0	0	0
tunnel	–5	0	0	0	–5	0	+28	0
inject	–5	–4	0	0	–9.3	0	0	0

synapse output. I left the *p*FET synapses turned on during tunneling, rather than turning them off like I did for the *n*FET-array experiments (see Table 2.1 and Table 2.2). I applied the voltages shown in Table 3.1 to read, tunnel, or inject synapse {1,1} selectively, while ideally leaving the other synapses unchanged.

3.1.4.1 Synapse Isolation

In the four-terminal *p*FET array, as in the four-terminal *n*FET array (see Section 2.1.4), the tunneling and drain terminals of the synapse transistors connect within rows, but not within columns. Consequently, the tunneling and IIHEI crosstalk between row synapses decreases exponentially with the voltage differential between the row synapses' floating gates, and the crosstalk between column synapses is negligible. I used 5V control-gate inputs, thereby achieving about a 4V differential between the floating gates of the selected and deselected row synapses; the resulting crosstalk between row synapses was <0.01 % for all operations.

I show synapse-isolation data in Figure 3.6. To obtain the data in part A, I first initialized all four synapses to I_s=100pA. I injected the {1,1} synapse up to 100nA, and then tunneled it back down to 100pA, while I measured the source currents of the other three synapses. As I expected, the row 2 synapses were unaffected by either the tunneling or the injection. Coupling to the {1,2} synapse also was small.

To obtain the data in part B of Figure 3.6, I first initialized all four synapses to I_s=100nA. I injected the {1,1} synapse down to 100pA, and then tunneled it back up to 100nA. As in the experiment of part A, crosstalk to the other synapses was negligible.

When I injected the {1,1} synapse, I applied –4V, rather than 0V, to the {1,2} synapse's control gate. I did so because hot-electron injection can occur in a *p*-type MOSFET by a mechanism different from that described in Section 3.1.2. If the floating-gate voltage exceeds the well

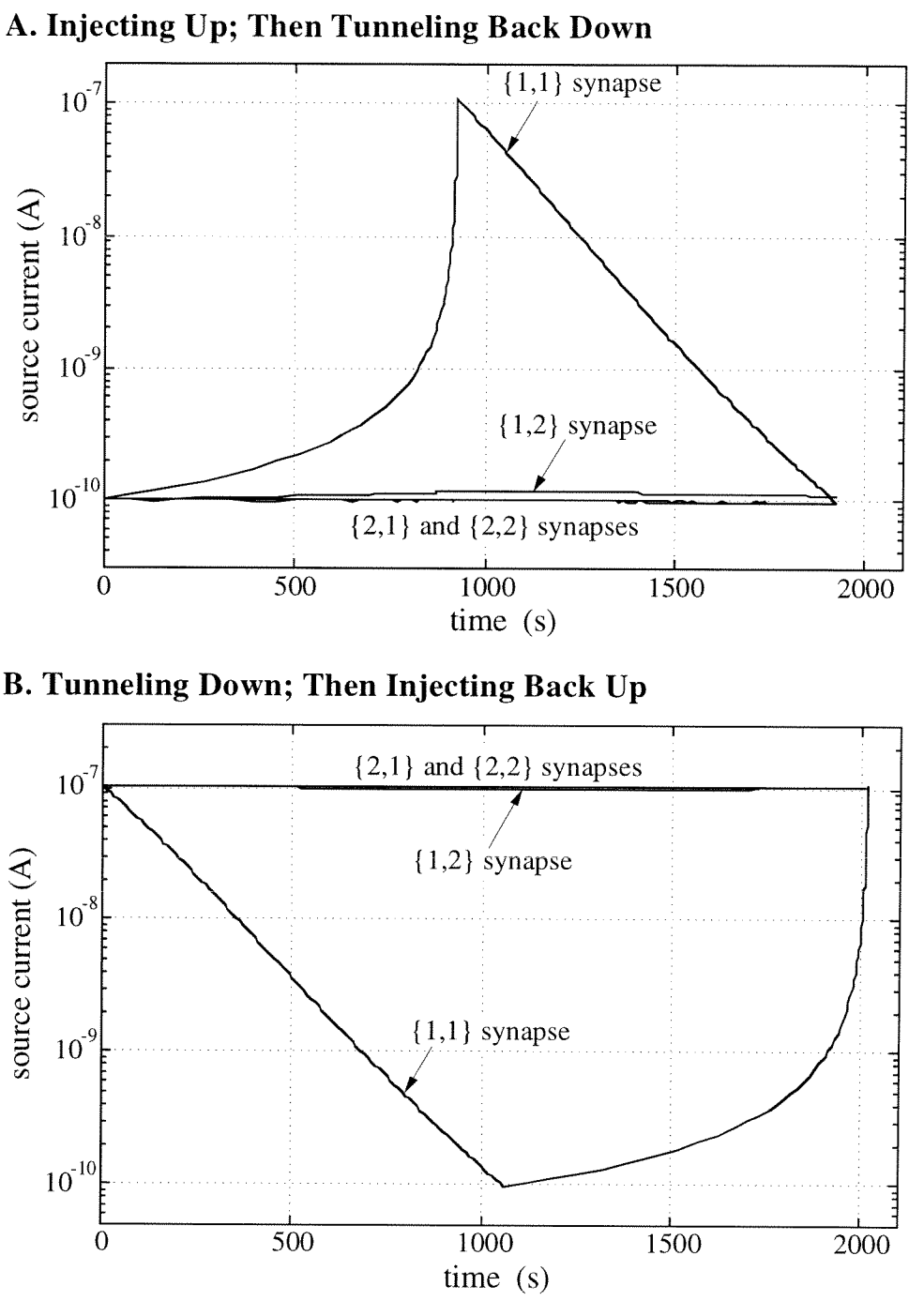


Figure 3.6 Isolation in a 2×2 array of four-terminal *p*FET synapses. The terminal voltages for both experiments are shown in Table 3.1 (see pg. 60). (A) I first injected the {1,1} synapse up to 100nA, then I tunneled it back down to 100pA, while I measured the source currents of the other three synapses. Crosstalk to the {1,2} synapse, defined as the fractional change in the {1.2} synapse's source current divided by the fractional change in the {1,1} synapse's source current, was 0.016% during injection, and was 0.007% during tunneling. (B) I first tunneled the {1,1} synapse down to 100pA, then I injected it back up to 100nA. Crosstalk to the {1,2} synapse was 0.004% during tunneling, and was 0.005% during injection.

voltage, and the drain-to-channel potential is large, electrons can inject onto the floating gate by means of a nondestructive avalanche-breakdown phenomenon at the MOS surface. I discuss this process in more detail in Section 3.1.6.

3.1.4.2 Synapse Weight Updates

A *p*FET synapse's weight updates derive from the tunneling and IIHEI oxide currents that alter the floating-gate charge. The weight-update rate, $\partial W/\partial t$, varies with the synapse's terminal voltages, which are imposed on the device, and with the source current, which is the synapse output. I repeated the experiment of Figure 3.6 (A), for several tunneling and injection voltages; in Figure 3.7, I plot the magnitude of the temporal derivative of the source current versus the source current, for a synapse transistor with (part A) a set of fixed tunneling voltages, and (part B) a set of fixed drain voltages. In both experiments, I held the control-gate input V_{in} fixed; consequently, these data show the synapse weight updates $\partial W/\partial t$, as can be seen by differentiating Eqn. (1.2). I now derive a weight-update rule that fits these data.

3.1.4.2.1 The Tunneling Weight-Decrement Rule

Tunneling weight updates in the four-terminal *p*FET synapse are functionally identical to tunneling weight updates in the *n*FET synapses, but with a sign inversion, because tunneling decreases, rather than increases, the synapse weight *W*. Consequently, the subthreshold weight-decrement rule is

$$\frac{\partial W}{\partial t} \approx -\frac{1}{\tau_{tun}} W^{(1-\sigma)} \tag{3.3}$$

where

$$\sigma \equiv \frac{V_f U_t}{\kappa V_{tun}^2} \tag{2.8}$$

and

$$\tau_{tun} \equiv \frac{Q_T}{I_{to}} e^{\frac{V_f}{V_{tun}}} \tag{2.9}$$

In the *n*FET synapses, the floating-gate–to–channel coupling coefficient, κ, is about 0.2. In the *p*FET synapse, κ is about 0.7. The difference is a consequence of the *n*FET synapses' additional channel doping. As a result, σ is smaller for a *p*FET synapse than it is for an *n*FET synapse.

3.1.4.2.2 The IIHEI Weight-Increment Rule

I now show that the IIHEI-induced weight increments follow a power law. I begin with the IIHEI gate current I_g:

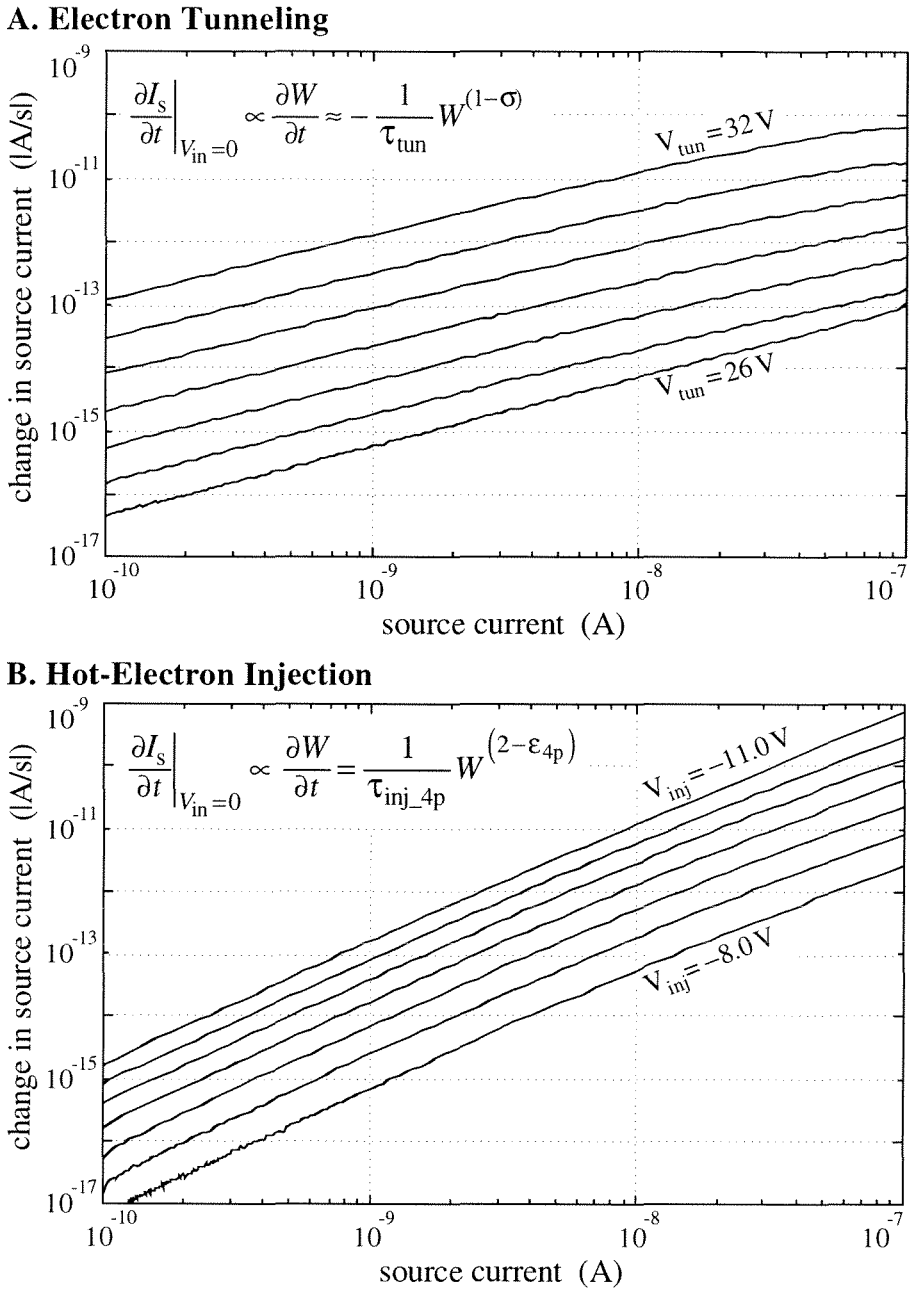


Figure 3.7 Four-terminal *p*FET-synapse (A) tunneling and (B) IIHEI weight updates. In both experiments, I measured the synapse's source current I_s versus time, and plotted $|\partial I_s/\partial t|$ versus I_s. I fixed the synapse's well-referenced terminal voltages; consequently, the change in I_s is a result of changes in the synapse's weight W. In (A) I applied $V_{in}=-5V$, $V_s=0V$, $V_{ds}=-4V$, and stepped V_{tun} from 26V to 32V in 1V increments; in (B) I applied $V_{in}=-5V$, $V_s=0V$, $V_{tun}=8V$, and stepped V_{ds} from $-8V$ to $-11V$ in $-0.5V$ increments. I turned off the tunneling and CHEI at regular intervals, to measure I_s. Because, for a fixed V_{in}, the synapse weight updates $\partial W/\partial t$ are proportional to $\partial I_s/\partial t$ (see Eqn. (1.2)), these data show that the weight updates follow a power law. The mean values of σ and ε_{4p} are 0.01 and 0.11, respectively.

$$I_g = \beta I_s e^{-\left(\frac{V_\beta}{V_{dc}+V_\eta}\right)^2} \tag{3.1}$$

Eqn. (2.13) describes, for a subthreshold MOSFET, the drain-to-channel potential, V_{dc}, in terms of V_{ds} and I_s:

$$V_{dc} = V_d - \Psi = V_{ds} - \Psi_o - U_t \ln\left(\frac{I_s}{I_o}\right) \tag{2.13}$$

I substitute V_{dc} into Eqn.(3.1):

$$I_g = \beta I_s e^{-\left(\frac{V_\beta}{V_{ds}+V_\eta-\Psi_o-U_t\ln\left(\frac{I_s}{I_o}\right)}\right)^2} = \beta I_s e^{-\left(\frac{V_\beta}{V_{ds}+V_\eta-\Psi_o}\right)^2\left[1-\frac{U_t}{V_{ds}+V_\eta-\Psi_o}\ln\left(\frac{I_s}{I_o}\right)\right]^{-2}} \tag{3.4}$$

I expand the exponent using $(1-x)^{-2} \approx 1+2x$, substitute for I_s using Eqn. (1.2), and solve:

$$I_g \approx \beta I_o e^{\left[\frac{(1-\varepsilon_{4p})\kappa' V_{in}}{U_t} - \left(\frac{V_\beta}{V_{ds}+V_\eta-\Psi_o}\right)^2\right]} W^{(1-\varepsilon_{4p})} \tag{3.5}$$

where

$$\varepsilon_{4p} \equiv \frac{2U_t V_\beta^2}{(V_{ds}+V_\eta-\Psi_o)^3} \tag{3.6}$$

I substitute Eqn. (3.5) into $\partial W/\partial t$ from Eqn. (2.4),

$$\frac{\partial W}{\partial t} = \frac{\beta I_o}{Q_T} e^{\left[\frac{(1-\varepsilon_{4p})\kappa' V_{in}}{U_t} - \left(\frac{V_\beta}{V_{ds}+V_\eta-\Psi_o}\right)^2\right]} W^{(2-\varepsilon_{4p})} \tag{3.7}$$

Now I define

$$\tau_{inj\_4p} \equiv \frac{Q_T}{\beta I_o} e^{\left[\left(\frac{V_\beta}{V_{ds}+V_\eta-\Psi_o}\right)^2 - \frac{(1-\varepsilon_{4p})\kappa' V_{in}}{U_t}\right]} \tag{3.8}$$

Finally, I substitute τ_{inj_4p} back into Eqn. (3.7), to get the IIHEI weight-increment rule:

$$\frac{\partial W}{\partial t} = \frac{1}{\tau_{inj\_4p}} W^{(2-\varepsilon_{4p})} \tag{3.9}$$

The parameter ε_{4p} varies with V_{ds}; the parameter τ_{inj_4p} varies with V_{ds} and with V_{in}.

3.1.4.2.3 The Synapse Weight-Update Rule

I obtain the complete weight-update rule, for the four-terminal *p*FET synapse, by adding Eqns. (3.3) and (3.9):

$$\frac{\partial W}{\partial t} \approx \frac{1}{\tau_{inj\_4p}} W^{(2-\varepsilon_{4p})} - \frac{1}{\tau_{tun}} W^{(1-\sigma)} \tag{3.10}$$

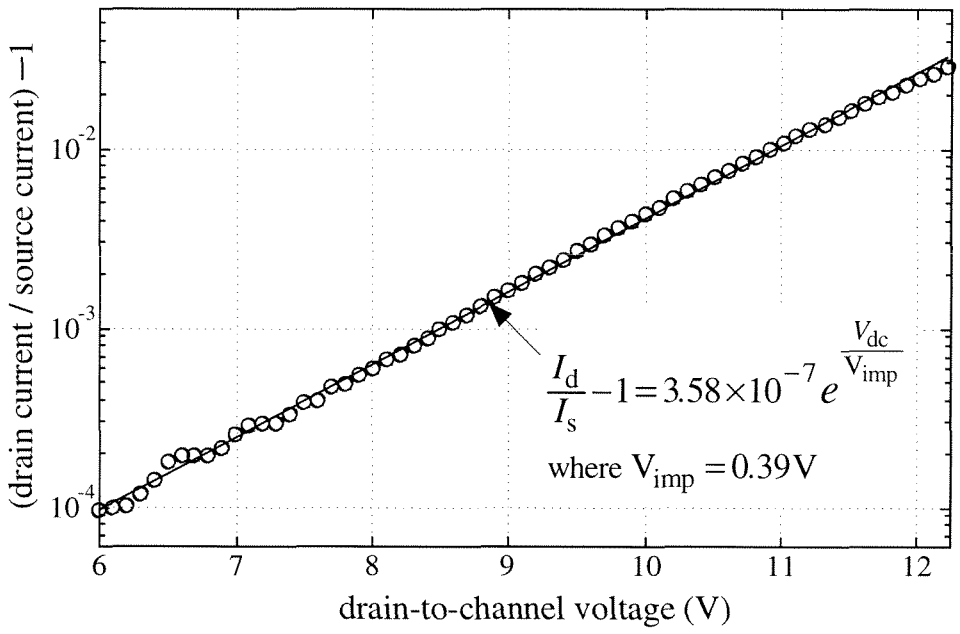


Figure 3.8 Four-terminal *p*FET-synapse impact ionization versus drain-to-channel voltage. I fixed the gate-to-channel voltage at $V_{gc} = 1.05\,V$ and the source current at $I_s = 10\,nA$, and measured the drain current I_d versus the drain-to-channel voltage V_{dc}. By plotting the data as efficiency (drain current I_d divided by source current I_s, minus one), I show the impact-ionization probability as a function of the drain-to-channel voltage. Although I can fit these data more carefully by using a modified lucky-electron model, the exponential fit is simpler, and models the data reasonably well over the entire drain-voltage range. Impact ionization in the four-terminal *p*FET synapse is markedly less efficient than it is in the four-terminal *n*FET synapse (see Figure 2.9), for two reasons. First, as a consequence of the bulk *p*-type implant that I add to the *n*FET transistor's channel region, the *n*FET synapse experiences a higher drain-to-channel electric field than does the *p*FET, thereby increasing the ionization likelihood. Second, the impact-ionization process is naturally more efficient for electrons (the *n*FET charge carriers) than it is for holes (the *p*FET charge carriers).

3.1.5 Impact Ionization Increases the Drain Current

I generate the electrons for oxide injection by means of hole-impact ionization in the four-terminal *p*FET synapse's drain-to-channel depletion region. As a result, the drain current exceeds the source current during IIHEI. I show hole impact-ionization data in Figure 3.8. Like I did for the four-terminal *n*FET synapse (see Section 2.1.5), I use a simple exponential fit, rather than a lucky-electron fit, to model these data. If I choose drain current, rather than source current, as the synapse output, I can rewrite the gate-current equation by replacing I_d with I_s according to my fit equation:

$$I_d = I_s\left(1+\gamma e^{\frac{V_{dc}}{V_{imp}}}\right) \tag{3.11}$$

where γ and V_{imp} are measurable fit constants.

3.1.6 An Alternate Injection Mechanism

Electrons can inject in a floating-gate *p*FET by means of a surface-induced avalanche-injection mechanism different from the IIHEI process described in Section 3.1.2. Frohman-Bentchkowsky used this alternate injection mechanism to write the first floating-gate MOS memory device [2]. In the 2μm *n*-well process that I use, the injection occurs as follows: If I raise the floating-gate voltage above the well voltage, then the *n*-type MOS surface accumulates electrons, and the depletion region separating the *n*-type surface from the *p*-type drain narrows. If I simultaneously apply a large negative drain voltage, then avalanche breakdown at the drain–surface *pn* junction liberates electron–hole pairs. The electrons are expelled from the drain by the drain-to-surface field, and can inject onto the floating gate. The avalanche-breakdown process is self limiting: The liberated electrons reduce the surface potential, decreasing the field and inhibiting further carrier generation.

When I turn off a *p*FET synapse, if I raise the control-gate voltage high enough to cause the floating gate to accumulate the MOS surface, and I simultaneously apply a large negative drain voltage, then electrons will inject onto the floating gate. Unfortunately, because the electron source is avalanche breakdown, I cannot control the magnitude of the injected charge accurately. Consequently, this injection mechanism probably is not useful for learning. In a learning array, however, avalanche injection can reduce the isolation between synapses. In Figure 3.9, I show the change in source current, versus time, for a *p*FET synapse transistor with five different control-gate (off) voltages. The larger the control-gate voltage I use to turn off the synapse, the greater the magnitude of this undesired gate current. To eliminate the effect, I use 1V control-gate signals to turn off the synapses in my *p*FET arrays (see Section 3.1.4.1).

3.2 A Guarded-*p*FET Synapse

The four-terminal *p*FET synapse's layout is large, because the transistor and the tunneling implant occupy separate wells. I now describe how I combine these two wells to obtain a more compact layout. I have named the new device a guarded-*p*FET synapse. A guarded-*p*FET synapse still is a four-terminal *p*FET synapse, and it still uses FN tunneling and IIHEI to modify the floating-gate charge. Consequently, I do not show the weight-update and array behavior for this synapse, because the data are nearly identical to those from Section 3.1. Instead, after I describe

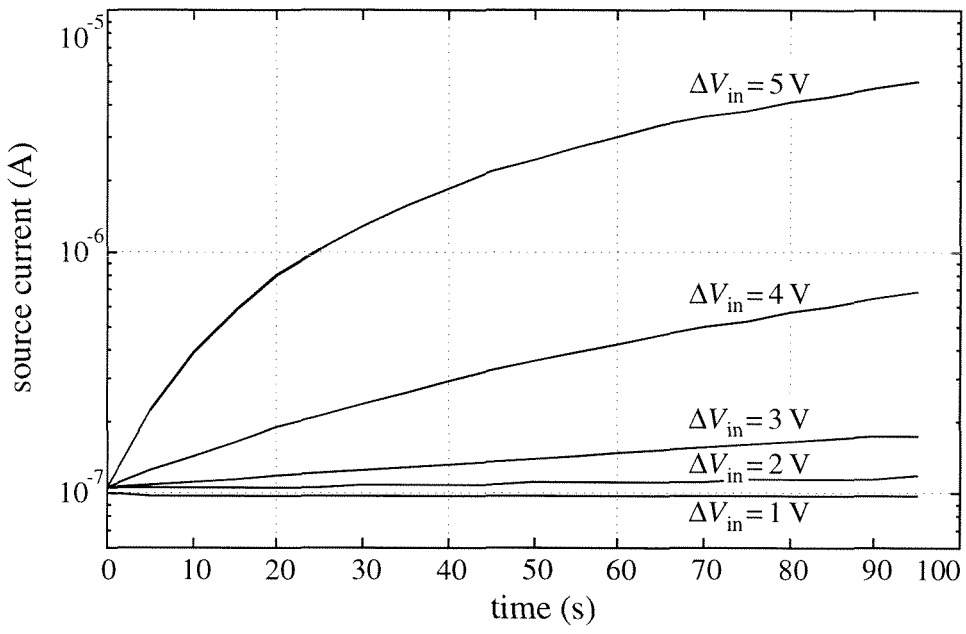


Figure 3.9 Four-terminal *p*FET-synapse avalanche injection versus control-gate voltage. For each of five experiments, I initialized the synapse's source current to $I_s=100\,nA$; fixed the well-referenced source voltage at $V_s=0\,V$; turned off the synapse by raising the control-gate input, V_{in}, by the amount indicated; and increased the drain-to-source voltage, V_{ds}, from 4 V to 10 V. At five-second intervals, I lowered V_{in}, decreased V_{ds} back to 4 V, and measured the synapse's source current. I plotted the source current versus the amount of time the synapse was turned off, for each of the five control-gate voltages. When $\Delta V_{in}>1\,V$, the synapse underwent avalanche injection as follows: When the floating-gate voltage was more positive than the well voltage, the *n*-type MOS channel accumulated electrons. When I increased V_{ds}, *pn* breakdown occurred at the junction between the accumulated *n*-type channel and the heavily doped p^+ drain. Electrons liberated by the breakdown process were expelled from the drain-to-channel depletion region at high energies, and injected onto the floating gate. I confirmed experimentally that the avalanche gate current increased exponentially with V_{ds}, consistent with my analysis of the injection process. To avoid this avalanche-induced gate current, I used 1 V control-gate pulses to turn off the *p*FET synapses in the array experiments (see Section 3.1.4.1).

the synapse's layout and operating concept, I describe how to use a guarded-*p*FET synapse as an analog EEPROM-type memory element.

I assume, for the moment, that to make the four-terminal *p*FET synapse smaller, I can simply merge the wells containing the tunneling implant and the floating-gate *p*FET. The tunneling implant comprises n^+ doped silicon, and an *n*-well comprises n^- doped silicon; consequently, a tunneling implant is a well contact. To induce electron tunneling, I must apply about 30 V across the gate oxide separating the floating gate and the well contact. I can do so by lowering the floating

gate by 30V; unfortunately, if the well potential is +12V, lowering V_{fg} by 30V requires using a large negative supply voltage on chip, and precludes reading the source current during tunneling. Alternately, I can raise the well potential by 30V, but doing so will cause *pn* breakdown at the reverse-biased drain-to-well and source-to-well *pn* junctions. To prevent this *pn* breakdown, I can raise the drain, source, and well potentials by 30V during tunneling, but then I cannot read the synapse's source current during tunneling. I employ a technique called junction guarding to solve this problem.

In a planar IC-fabrication technology, the implant-impurity concentrations usually are much higher near the semiconductor surface than they are in the bulk. Consequently, the electric field across a *pn* junction is highest at the surface, and reverse-bias junction breakdown usually occurs near the surface [3]. Junction guarding is a well-known technique for reducing the surface electric field. By surrounding an implant with a MOS guard ring, and applying the high voltage to both the implant and the ring, I widen the depletion region at the semiconductor surface, thereby decreasing the peak electric field and increasing the junction's breakdown voltage [3]. In Figure 3.10, I show *pn*-breakdown voltage, versus guard-ring voltage, for a heavily doped n^+ implant (in *p*-type substrate) surrounded by a polysilicon-gate guard ring. For junction voltages in the 30V range, the *pn*-breakdown voltage increases nearly one to one with the guard-ring voltage.

I show the guarded-*p*FET synapse in Figure 3.11. This synapse has three notable features: (1) the layout contains a single *n*-type well, (2) the floating gate abuts the n^+ well contact, and (3) the floating gate surrounds the *p*-type drain and source implants. I apply positive high voltages to the n^- well, tunnel electrons from the floating gate to the n^+ well contact, and use the floating gate to guard the drain and source implants against *pn* breakdown. From the well's perspective, the drain and source implants are at large negative voltages (although the voltages still are positive with respect to the substrate); consequently, for guarding, the floating gate must also be at a large negative voltage. In a subthreshold *p*FET, the floating-gate voltage will always be near the source voltage; consequently, the floating gate is naturally at the proper potential for guarding. Simply by surrounding the drain and source implants with the floating gate, I guard these junctions against *pn* breakdown during tunneling.

A guarded *p*FET remains a fully functional *p*-type MOSFET. The only differences between a guarded *p*FET and a conventional *p*FET are the larger well-voltage range, and larger drain-to-gate and source-to-gate overlap capacitances. Consequently, in a guarded-*p*FET synapse I can *simultaneously* (1) raise the well voltage, causing electron tunneling from the floating gate to the n^+ well contact; (2) adjust the floating gate and source voltages to effect subthreshold source currents; and (3) lower the drain voltage, causing IIHEI.

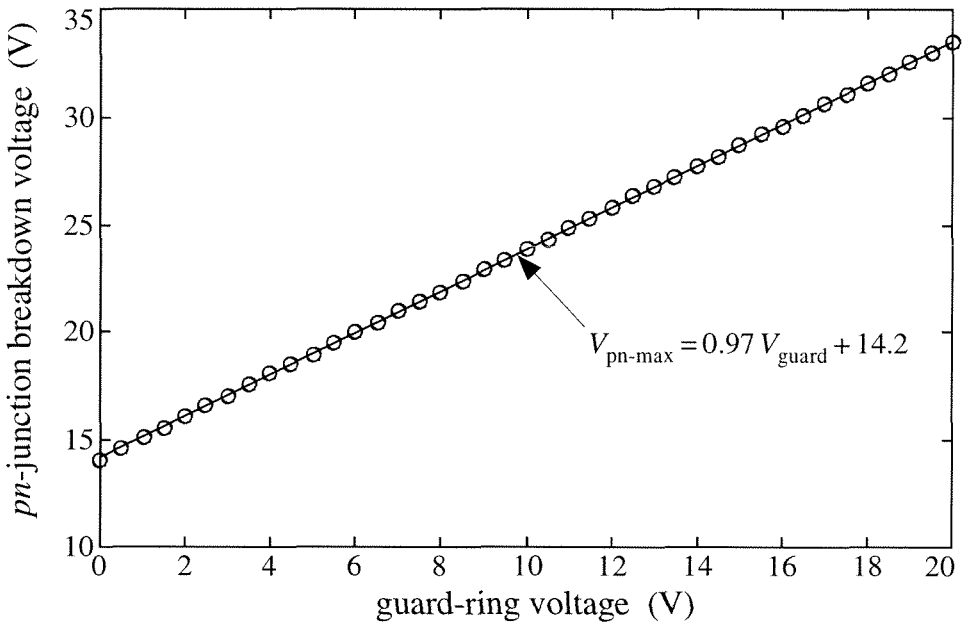


Figure 3.10 Junction-diode breakdown voltage versus guard-ring voltage. I fabricated an isolated n^+ implant, in the p^- substrate, and surrounded the n^+ implant with a polysilicon-gate guard ring (see Figure 3.11 for an example). I stepped the guard-ring voltage from 0V to 20V in 0.5V increments; for each guard-ring voltage, I ramped the n^+ voltage upward, starting from the substrate potential (ground), and ending when the leakage current from the n^+ implant to the p^- substrate just exceeded 10nA. I plotted the maximum voltage that I applied to the n^+ implant, $V_{pn\text{-}max}$, versus the guard-ring voltage. These data show clearly that guard rings extend the reverse-bias breakdown voltage of implanted *pn* junctions. Although I measured these data from an n^+ implant within substrate, a guarded p^+ implant within an n^- well behaves similarly.

3.2.1 Electron Tunneling Decreases the Weight

I show the tunneling-junction layout, and an energy-band diagram for the tunneling process, in parts A and C of Figure 3.11, respectively. The layout and band diagram are identical to those for the four-terminal *p*FET synapse (compare Figure 3.11 with Figure 3.1). I apply positive high voltages to the n^+ well contact to cause electron tunneling, thereby decreasing the synapse weight W.

The guarded-*p*FET synapse's floating gate extends from the MOSFET, over the n^- well, to the n^+ well contact. As I did for the three-terminal *n*FET synapse (see Figure 2.11), I placed a field-oxide channel stop beneath the floating-gate extension, ostensibly to prevent the channel–surface depletion layer from reaching the n^+ well contact. Unfortunately, as I have already described for the three-terminal *n*FET synapse (see Section 2.2.6), when the well voltage exceeds

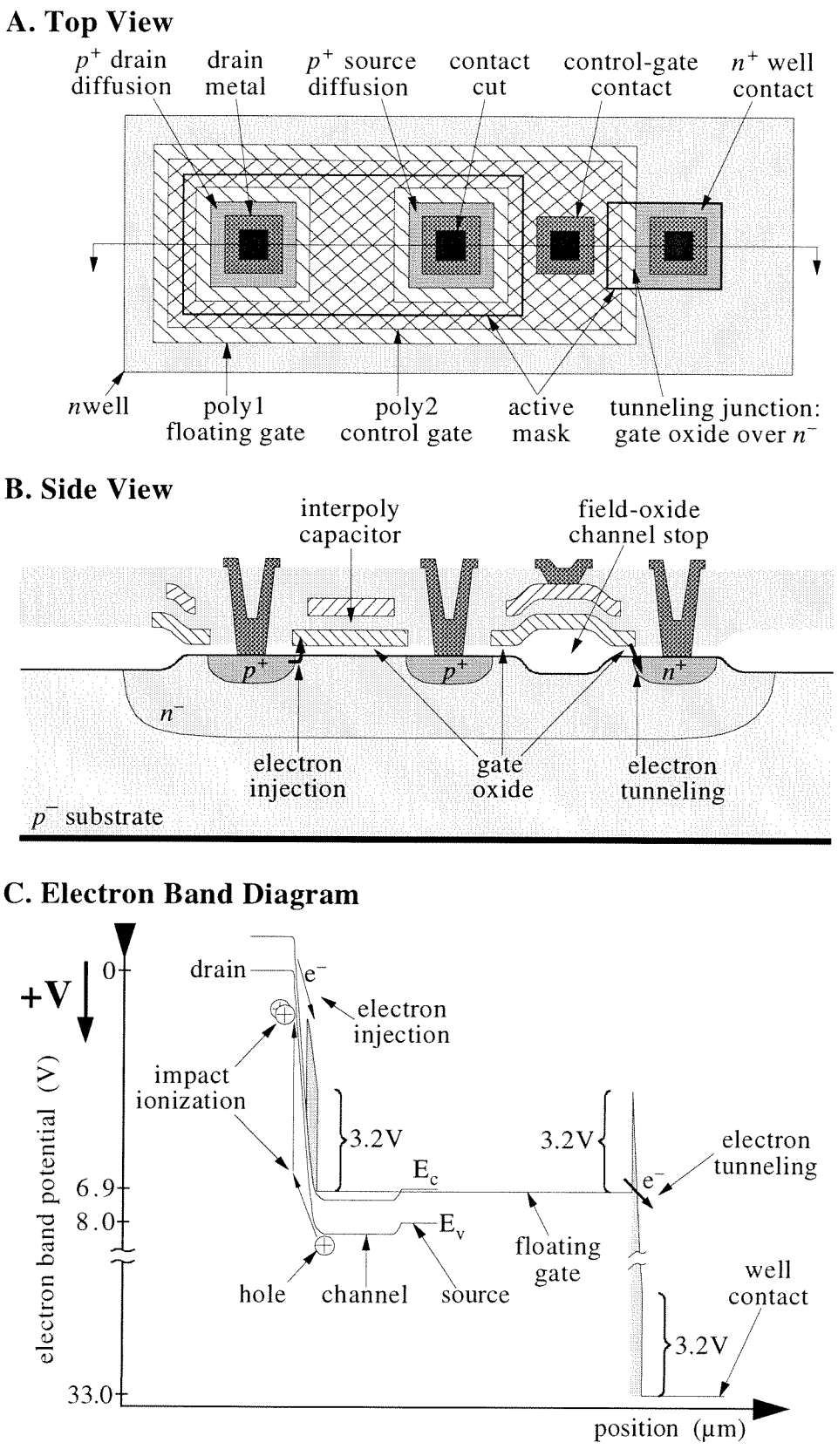


Figure 3.11 The guarded-*p*FET synapse, showing the electron tunneling and injection locations. As I did in

Figure 3.1, I have aligned the three diagrams vertically, drawn diagrams A and C to scale, exaggerated the vertical scale in diagram B, referenced the voltages in the band diagram to the source potential, drawn the gate-oxide band diagram in the channel direction, and assumed subthreshold source currents (I_s<100nA). In the 2μm process, the synapse length is 48μm, and the width is 18μm. In a learning array, I can place multiple synapses within a single *n*-type well; consequently, the effective synapse dimensions are smaller (about 38μm×14μm, depending on the layout). The floating-gate surrounds the p^+ drain and source implants, guarding these junctions against *pn* breakdown during well tunneling. I draw the active mask around the p^+ implants, thereby ensuring that the floating-gate guard rings subtend gate oxide, rather than field oxide, near the p^+. I orient the transistor so that the drain implant is farther from the well contact than is the source implant, because the well-contact–to–drain voltages are larger than the well-contact–to–source voltages. The field-oxide channel stop was intended to prevent the channel-surface depletion layer from reaching the tunneling junction. Unfortunately, during tunneling, a leakage current flows from the n^+ well contact to the *p*-type channel (see the discussion of tunneling-junction leakage in Section 2.2.6). To eliminate this leakage path in the test device, I routed the floating gate from the channel region to the well contact in metal, rather than in polysilicon. I expect this leakage problem to disappear when I use a more modern process with lower tunneling voltages. With a 115fF interpoly capacitor as shown, the coupling coefficient between the poly2 control gate and the poly1 floating gate is about 0.35. I enlarged the interpoly capacitor to 1pF in the test device, thereby increasing the coupling coefficient to 0.8. I decrease the synapse weight *W* by tunneling electrons to the well contact; I increase the weight by means of IIHEI in the channel-to-drain depletion region.

the floating-gate voltage by about 30V, avalanche breakdown occurs at the n^+ tunneling implant. In the guarded-*p*FET synapse, a leakage current flows during tunneling from the n^+ well contact, beneath the channel stop, to the *p*-type source and drain.

In Section 3.2.4, I describe an alternate guarded-*p*FET synapse that does not exhibit junction leakage during tunneling. For the synapse of Figure 3.11, in the 2μm process, I modify the layout by routing the floating-gate extension in metal, rather than in polysilicon, thereby eliminating the leakage current entirely (but also, unfortunately, enlarging the layout).

3.2.2 Electron Injection Increases the Weight

I increase the synapse weight *W* by injecting electrons onto the floating gate. The IIHEI process in the guarded-*p*FET synapse is identical to the IIHEI process in the four-terminal *p*FET synapse (compare Figure 3.11 with Figure 3.1). In Figure 3.12, I show the guarded-*p*FET synapse's IIHEI efficiency (gate current I_g divided by source current I_s) versus both the drain-to-channel and the well-contact–to–channel potentials. As I did for the four-terminal *p*FET synapse

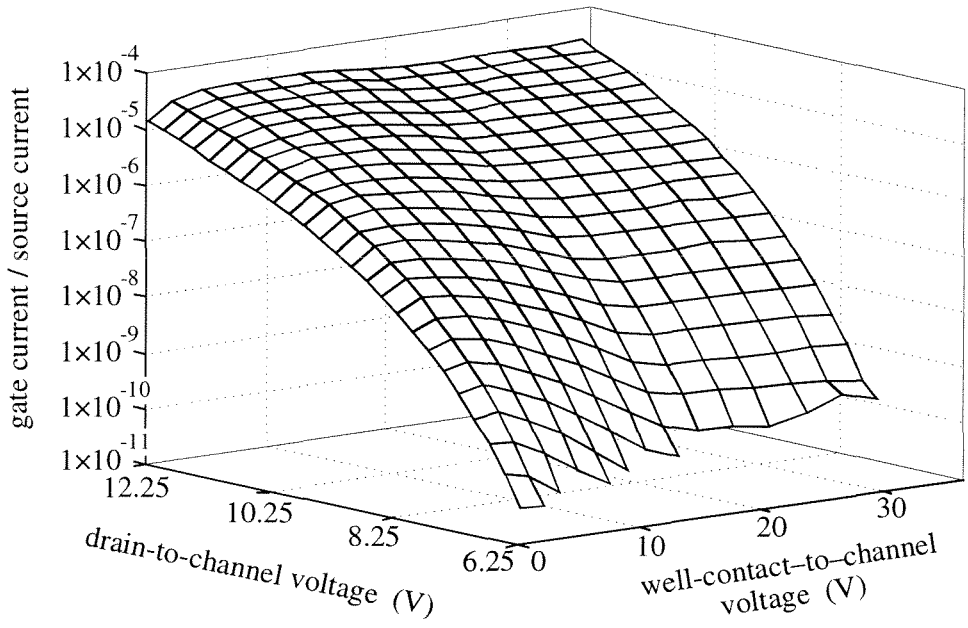


Figure 3.12 Guarded-*p*FET synapse IIHEI surface plot. I measured the IIHEI gate current I_g versus both the well-contact–to–channel potential, V_{wc}, and the drain-to-channel potential, V_{dc}, for a fixed source current I_s=10nA and a fixed floating-gate voltage V_{fg}=2V (relative to substrate). I plotted the gate current I_g divided by the source current I_s. In the subthreshold regime, I_g increases linearly with I_s; consequently, these data show the IIHEI efficiency for the entire subthreshold source-current range. I used V_{fg}=2V, rather than the more typical V_{fg}=10V, to avoid field-oxide electron injection where the floating gate exits the high-voltage well (I route the floating gate from the well to measurement circuitry in substrate). In most learning applications, the floating gate will not exit the well. The dependency of the IIHEI efficiency on V_{dc} is similar to that for the four-terminal *p*FET synapse (see Figure 3.2). The dependency of the IIHEI efficiency on V_{gc} is different from that for the four-terminal *p*FET synapse, because the well pinches off at roughly V_{wc}=18V. I therefore plot these data versus V_{wc}, rather than versus V_{gc}, because, for V_{wc}>18V, changes in the well-contact voltage do not cause commensurate changes in V_{gc}. The data of Figure 3.13 show the effect more clearly. If I maintain V_{wc}>18V, then the guarded-*p*FET synapse's IIHEI efficiency becomes, for all practical purposes, independent of the well-contact voltage.

(see Figure 3.2), I plot the data as efficiency because the gate current increases linearly with the source current over the subthreshold source-current range, and I reference the drain potential to the channel potential because the hot-electron population derives from the drain-to-channel electric field. However, unlike in Figure 3.2, I plot the guarded-*p*FET efficiency data versus the well-contact–to–channel potential, rather than versus the gate-to-channel potential, to illustrate an interesting phenomenon: The IIHEI efficiency does not increase monotonically with the well-contact voltage. The data of Figure 3.13 show the effect clearly.

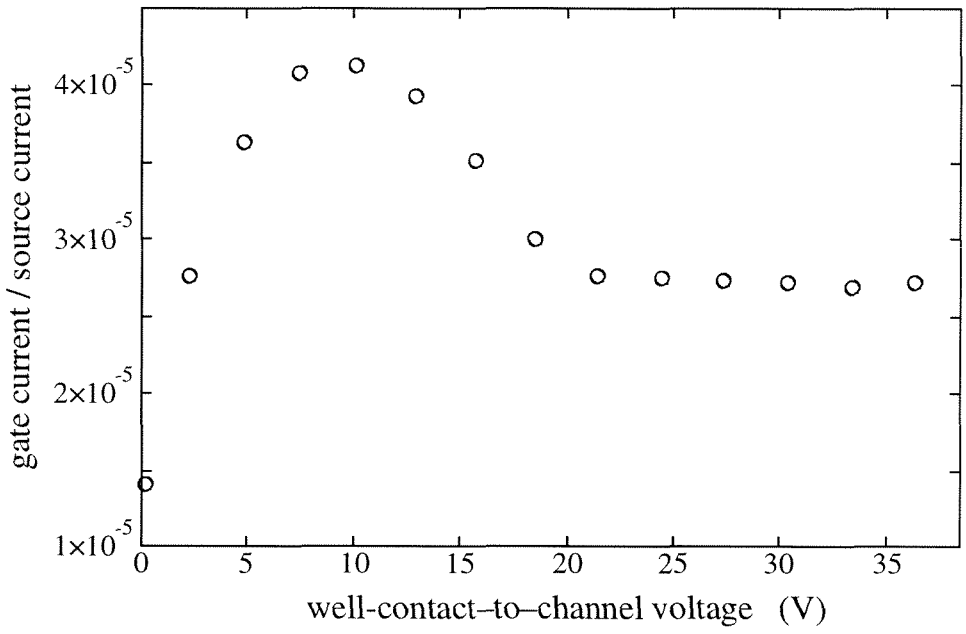


Figure 3.13 Guarded-*p*FET synapse IIHEI efficiency versus well-contact voltage. I replotted a subset of the data from Figure 3.12, for a single fixed drain-to-channel voltage $V_{dc}=12.25\,V$. For $V_{wc}<10\,V$, the IIHEI efficiency increased with V_{wc}, for the following reason: When I increased V_{wc}, I simultaneously increased V_{gc}, to maintain a fixed source current (see the discussion in Section 3.2.2). Consequently, these data are consistent with the data in Figure 3.2. For $10\,V<V_{wc}<18\,V$, the IIHEI efficiency decreased with increasing V_{wc}, because the drain guard ring widened the drain-to-channel depletion region, decreasing the drain-to-channel electric field and with it the hot-electron population. For $V_{wc}>18\,V$, the IIHEI efficiency became independent of V_{wc}, as a result of well pinchoff (see Figure 3.14 and Figure 3.15). I estimated the well-pinchoff voltage to be about 18V not so much from these data, but rather from the data in part B of Figure 3.15 (with a source voltage $V_s=12\,V$).

For well-contact–to–channel voltages less than about 10V, the IIHEI efficiency increases with the well-contact voltage. I can view these data another way: As I raise the well-contact voltage, I must simultaneously lower the floating-gate voltage (increase the magnitude of the floating-gate voltage relative to the source), to maintain a fixed channel potential and thereby a fixed channel current. These data therefore show that a guarded *p*FET's IIHEI efficiency increases with gate-to-channel voltage, just like in the four-terminal *p*FET synapse (see Figure 3.2).

For well-contact–to–channel voltages greater than about 10V, but less than about 18V, the IIHEI efficiency decreases with increasing well-contact voltage. I believe that this decrease happens for the same reason that the floating-gate guard rings inhibit drain-to-well junction breakdown: The potential difference between the well and the guard ring widens the surface depletion region near the p^+ drain implant, decreasing the peak electric field. Guard rings also widen the

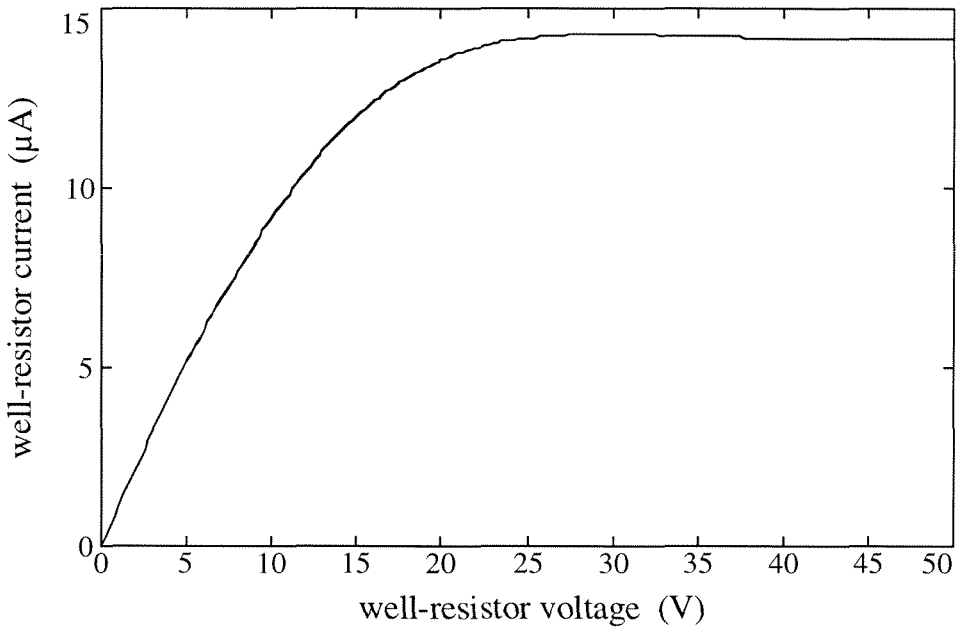


Figure 3.14 Well-resistor pinchoff. I fabricated, in substrate, a 1900μm×10μm strip of n^- well (a well resistor) with n^+ contacts at each end, and a polysilicon cap (poly1 over field oxide) over most of the length. I grounded the substrate, the poly1 cap, and one of the n^+ contacts, and I swept the other n^+ contact from 0V to 50V while measuring the well-resistor current. For low voltages, the well looked like a 1MΩ resistor. As I increased the voltage, the well-to-substrate and well-to-channel depletion regions widened, decreasing the amount of undepleted n^- available for current flow and thereby increasing the resistance. At about 20V, the depletion regions met, pinching off the well in a manner analogous to JFET channel pinchoff [3], causing the resistance$\rightarrow\infty$. For a well resistor without the poly1 cap, the well pinches off at 40V to 50V, rather than at 20V, because the well-to-substrate depletion region must widen all the way to the silicon surface.

drain-to-channel depletion region, decreasing both the electric field and the IIHEI efficiency.

For well-contact–to–channel voltages greater than about 18V, the IIHEI efficiency is independent of the well-contact voltage. This effect is intriguing—indirect evidence, in particular the well-resistor data that I show in Figure 3.14, point to well pinchoff as the likely cause. As I increase the well-contact voltage, the well-to-substrate and the well-to-channel depletion regions widen: When the well-contact–to–channel voltage reaches 18V, the depletion regions meet, pinching off the well in a manner analogous to JFET channel pinchoff [3]. For all practical purposes, the well region beneath the *p*FET becomes semi-insulating; further increases in the well-contact potential do not affect the channel potential. If this hypothesis is correct, then well pinchoff may also be implicated in the tunneling-junction leakage problem (see Section 2.2.6), in ways that I do not yet understand.

3.2.3 Well Pinchoff Isolates the Tunneling Implant

Well pinchoff isolates the well-contact tunneling implant from the *p*-type MOSFET, thereby allowing separate and independent control of the gate-oxide tunneling and the MOS transistor operation. As I show in Figure 3.15, when the well-contact–to–source-voltage is equal to about 18V, both the floating-gate–to–channel coupling coefficient, κ, and the transistor's threshold voltage, V_t, become constant. By the data of Figure 3.13, further increases in the well-contact voltage do not affect the IIHEI efficiency; by the data of Figure 3.15, further increases do not affect the transistor's channel current. Consequently, well pinchoff decouples the tunneling implant from the *p*-type MOSFET, and the guarded-*p*FET synapse looks, for all practical purposes, just like the four-terminal *p*FET synapse. Of course, κ is higher, and as a result the parameter σ in the tunneling weight-update rule, Eqn. (3.3), is smaller, but these improvements are secondary. The primary result is that the guarded-*p*FET synapse operates in a fashion similar to the four-terminal *p*FET synapse, but with a more compact layout.

A typical operating condition for this synapse is V_s=12V, V_{well}=30V (both voltages are referenced to the substrate), and I_s=10nA. For these conditions, the floating-gate voltage is roughly 9V (the large gate-to-source voltage is a consequence of the threshold voltage increase—see part B of Figure 3.15), and the tunneling-oxide voltage is 21V. The well is pinched off, but the tunneling-oxide voltage is too small to cause measurable electron tunneling. If V_d>7V, then there is no measurable IIHEI, and the synapse's weight remains nonvolatile. If I raise the well-contact voltage to 40V, the well remains pinched off, and I do not alter either the transistor's source current or its IIHEI gate current directly (there is parasitic capacitive coupling from the tunneling implant to the floating gate, but this coupling is common to all the synapses). I do, however, cause measurable electron tunneling, decreasing the synapse's weight *W*. Alternately, if I lower the drain voltage to ground, then I induce IIHEI, increasing the synapse's weight *W*. Because well pinchoff decouples the tunneling implant from the *p*FET, electron tunneling and IIHEI can occur simultaneously.

3.2.4 An Alternate Tunneling Junction

During tunneling, field-oxide–induced junction breakdown at the n^+ well contact causes a leakage current to flow from the well contact to the *p*FET's drain and source. I therefore investigated alternate tunneling junctions where the floating gate does not abut the n^+. I first built a guarded-*p*FET synapse without a floating-gate extension to the n^+ well contact, in the hopes of inducing FN tunneling through the gate oxide that subtends the channel. Unfortunately, the *p*FET's source potential pins the MOS-channel potential; consequently, the channel is at or near

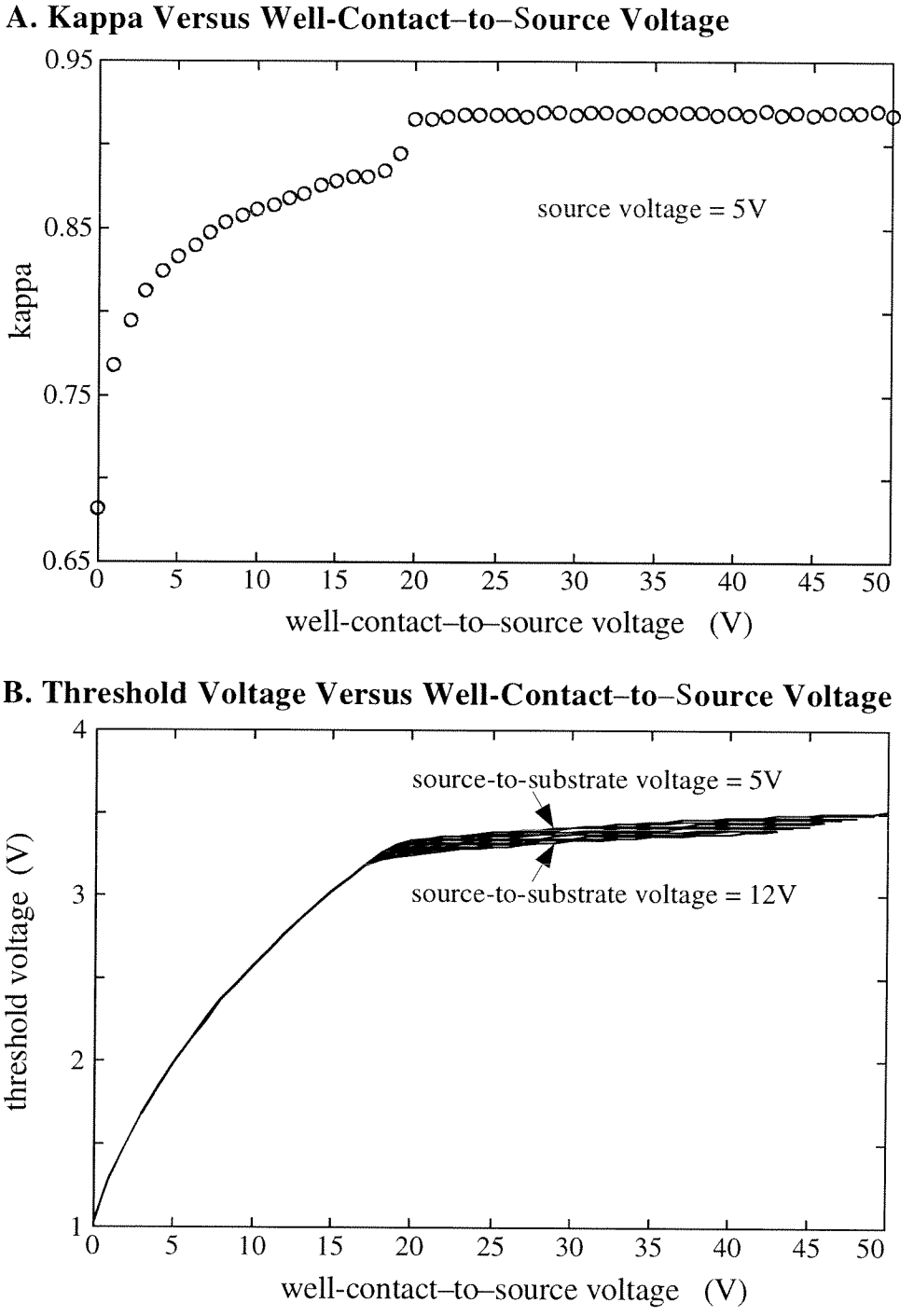


Figure 3.15 Guarded-*p*FET synapse kappa and threshold voltage versus well-contact voltage. For the synapse of Figure 3.11, I swept the well-contact voltage from the source voltage, V_s, to 55 V (substrate referenced), in 1 V steps. At each step, I measured (A) the gate-to-channel coupling coefficient, κ, and (B) the transistor's threshold voltage, V_t, and I plotted these data versus the well-contact–to–source voltage, V_{ws}. When V_{ws} exceeded about 18 V, well pinchoff depleted the n^- silicon beneath the transistor, causing the well

to become effectively semi-insulating. The data in (A) show that, when $V_{ws}>18\,V$, the transistor's backgate was eliminated, and κ became constant. I do not yet understand what caused the discontinuity in the data at $V_{ws}=18\,V$, and why κ did not reach unity. In (B), I measured V_t for seven values of V_s ranging from $V_s=5\,V$ to $V_s=12\,V$ (substrate referenced) in 1 V steps. The higher the source voltage, the higher the well-contact voltage, and the wider the initial well-to-substrate depletion region. Consequently, the well-contact–to–source voltage at which the well pinched off decreased as V_s increased.

the source voltage, rather than near the well-contact voltage, and the resulting oxide voltage is insufficient for tunneling. (Note: MOSFETs with 40 Å or thinner gate oxides can tunnel electrons directly from the floating gate to the channel—see the discussion of direct tunneling in Section 5.1.) To isolate the tunneling region from the *p*FET's source, I fabricated the guarded-*p*FET synapse shown in Figure 3.16. In this device, electrons tunnel from the floating gate to the n^- well through what I call a bowl-shaped tunneling junction.

I extend the *p*FET's floating gate over a region of field oxide, and I place an isolated, 4 μm×4 μm square bowl of gate oxide within this field oxide. The gate-oxide bowl has n^- silicon beneath it, the polysilicon floating gate above it, and field oxide on all four sides. I apply a high voltage to the n^- well, causing electrons to tunnel from the floating gate, through the gate-oxide bowl, to the n^-. The floating gate depletes the n^- silicon immediately beneath the bowl, causing a potential drop from the bulk n^- to the MOS surface. Consequently, bowl tunneling requires well voltages roughly 5 V higher than those required to tunnel at an n^+ well contact. However, because I tunnel through a gate-oxide surface, rather than at an edge, oxide trapping is reduced.

Bowl-shaped tunneling junctions do not exhibit the leakage currents that I observed at n^+ well-contact tunneling junctions (see Section 2.2.6), consistent with my analysis that, at an n^+ tunneling implant, *pn* breakdown occurs where the induced *p*-type surface abuts the n^+ well contact. Unfortunately, electron tunneling in a bowl-shaped junction also presents inconsistencies with my prior analysis. A parasitic channel should form beneath the floating-gate extension that separates the bowl-shaped tunneling junction from the *p*FET, causing the *p*FET's channel potential to pin the surface potential beneath the bowl, preventing tunneling. However, the bowl-shaped junctions do tunnel, indicating that I do not yet fully understand the formation and characteristics of parasitic field-oxide channels.

Although the bowl-shaped tunneling junction does eliminate the *pn*-breakdown problem, its turn-on delay (the delay between the application of a high well voltage and the onset of electron tunneling) is long. In Figure 3.17, I show the amount of charge tunneled through a bowl-shaped oxide, versus the amount of time the well voltage was pulsed high, for three different well-pulse

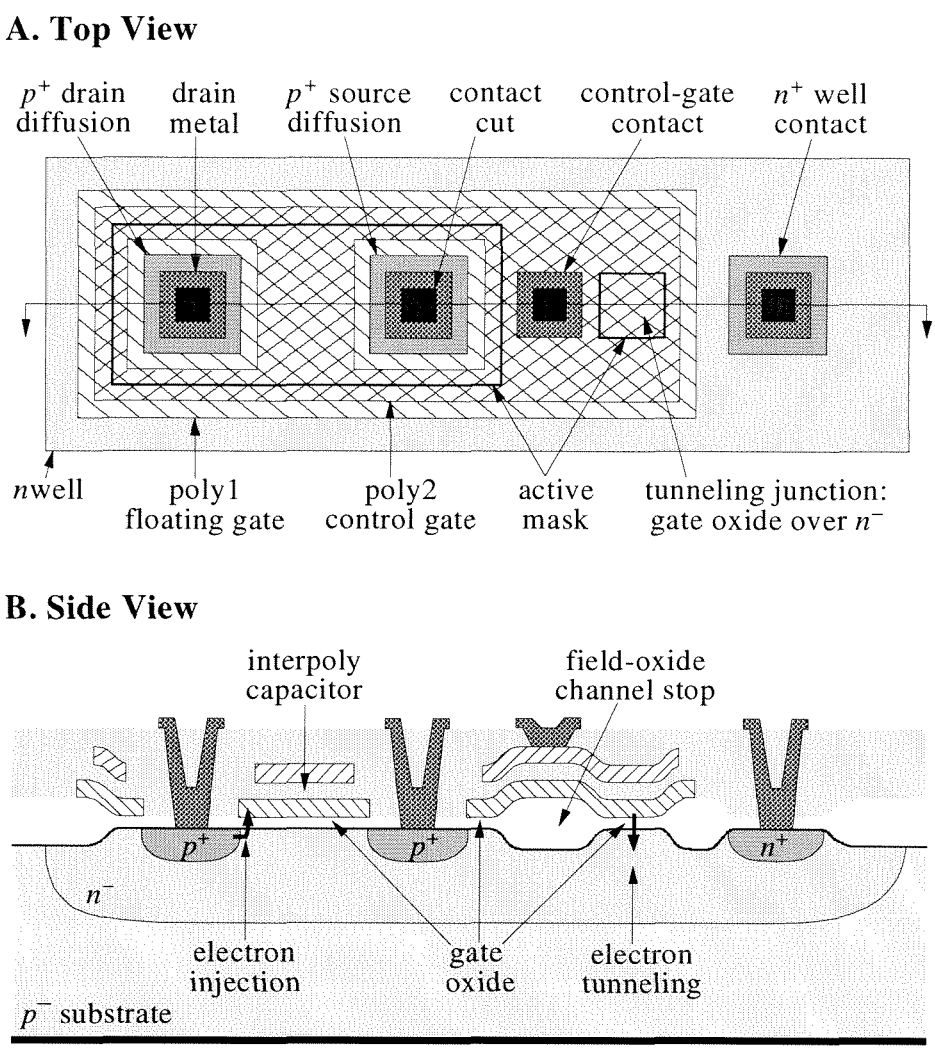


Figure 3.16 A guarded-*p*FET synapse without tunneling-junction leakage. I have aligned the two diagrams vertically, and I have exaggerated the vertical scale in diagram B. The only difference between this synapse and the one shown in Figure 3.11 is the tunneling-junction layout. In this synapse, I form a bowl-shaped tunneling junction using an isolated active mask, separated from both the transistor and the well contact. This synapse tunnels electrons from the floating gate to the n^- well; the synapse in Figure 3.11 tunnels electrons from the floating gate to the n^+ well contact. The band diagrams for the tunneling and IIHEI processes are identical for both devices; consequently, I do not redraw the band diagram here. Because the floating gate in this synapse does not abut the n^+ well contact, the channel-surface depletion layer does not reach the n^+, eliminating the leakage current shown in Figure 2.21 completely. In the 2μm Orbit process, the synapse length is 54μm, and the width is 18μm. Because I anticipate using multiple synapses within a single *n*-type well, the effective synapse dimensions are smaller (about 49μm×14μm, depending on the layout).

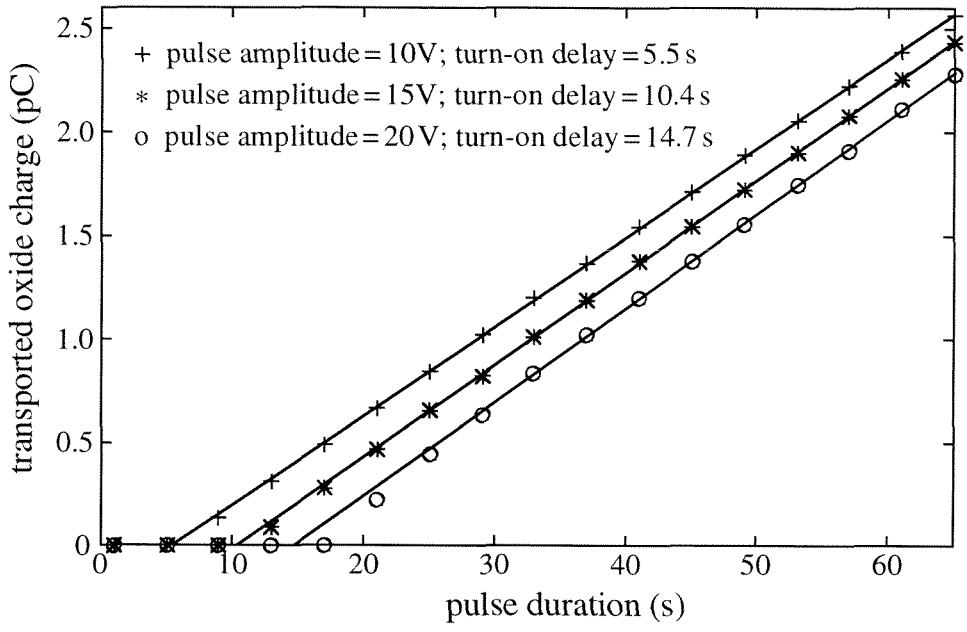


Figure 3.17 Bowl-shaped tunneling junction turn-on delay. For the synapse of Figure 3.16, I fixed the floating-gate, drain, and source voltages at 11V (substrate referenced), and I set the resting well voltage to 37V (+), 32V (*), and 27V (o). I pulsed the well to 47V for the time indicated, and I measured the amount of charge tunneled through the oxide versus the tunneling pulsewidth. The turn-on delay (the delay between the start of the tunneling pulse and the onset of electron tunneling) always exceeded several seconds, and increased with the well-pulse amplitude. The reason for the turn-on delay is as follows: The well voltage exceeded the floating-gate voltage, so the silicon surface beneath the bowl was depleted. The depletion-region depth increased with the voltage differential between the floating gate and the well. When I pulsed the well high, holes diffused to the silicon surface to deplete more of the n^- well, raising the surface potential. Unfortunately, the only source for these holes was thermal generation. As a result, the surface potential increased only gradually. Because electron tunneling increases exponentially with the tunneling-oxide voltage, no appreciable oxide current flowed until the surface was fully depleted. I can use this bowl-shaped tunneling junction in systems for which the tunneling voltage is a slowly varying analog quantity, but, because of the long turn-on delay, I cannot use it in systems in which I pulse–tunnel a synapse.

voltage amplitudes. The turn-on delay can exceed 10 seconds—an impracticably long time for a pulse-based learning system. The cause is the depletion region at the silicon surface beneath the bowl. As a result of the voltage differential between the floating gate and the n^- well, the surface region beneath the gate oxide is depleted, and the depletion-region depth varies with the voltage differential between the floating gate and the well. If I pulse the well high, I must provide holes to the silicon surface to widen this depletion region. Unfortunately, the only hole source is thermal-carrier generation. Consequently, the depletion region takes many seconds to widen. Al-

though I can use bowl-shaped tunneling junctions in systems for which the well-tunneling voltage is a slowly varying analog quantity, I cannot use them in systems in which I pulse–tunnel a synapse.

3.2.5 An Analog EEPROM with Self-Convergent Writes

I now describe how I use my guarded-*p*FET synapse as an analog-EEPROM transistor. There is a need for nonvolatile analog storage in standard CMOS processes [4, 5], to store bias voltages or currents, to record continuously valued analog signals [6], to store analog weights in silicon neural networks, and to permit multilevel digital memories. This need has not been satisfied adequately by commercial *n*FET EEPROMs, primarily because conventional EEPROM transistors do not permit simultaneous memory reading and writing. Most analog EEPROM implementations require iterative writes: The memory first is written, then is read; the written and read values then are compared, and the error is used to write a correction. This cycle is repeated until the error is within prescribed bounds.

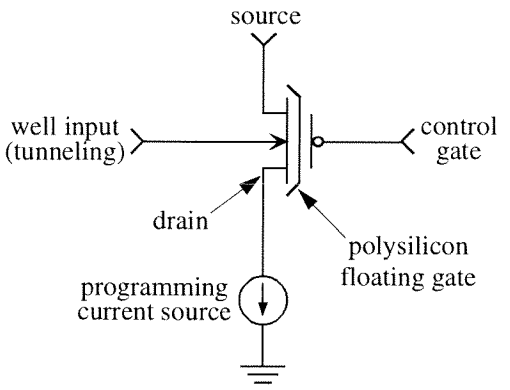
Unlike conventional EEPROM transistors, my guarded-*p*FET synapse allows simultaneous memory reading and writing. Consequently, I can apply continuous negative feedback during the write process, to store an analog memory value in a single-step write. This process is called self-convergent writing: An intrinsic, self-limiting feedback path within the transistor itself ensures that the analog memory value is stored accurately.

3.2.5.1 Writing the Memory

I fabricated a prototype analog EEPROM, to investigate the array write and erase procedures, and the memory-write accuracy. Each memory cell comprised a single guarded-*p*FET synapse. I show the write process in part A of Figure 3.18, and the array in part B of Figure 3.18. I always erased a cell before writing. I describe the erase process in Section 3.2.5.2.

I chose to read and write drain-current values. To write a cell, I applied a low voltage to the row control-gate input, and a programming sink current to the column-drain wire. I generated the cell's programming sink current from a current source, as shown in part A of Figure 3.18. As long as this programming current exceeded the cell's drain current, the drain voltage remained low, and electrons injected onto the cell's floating gate. Electron injection caused the floating-gate–to–source voltage to increase, thereby increasing the cell's drain current. As soon as the drain current exceeded the programming current, the drain voltage rose, turning off the injection. IIHEI closed a negative-feedback loop around the inverting amplifier formed from the guarded-*p*FET synapse and the programming current source [7]. This loop adapted the cell's floating-gate charge to equalize the programming and *p*FET-drain currents.

A. The Self-Convergent Write Process



B. Writing An Analog Array

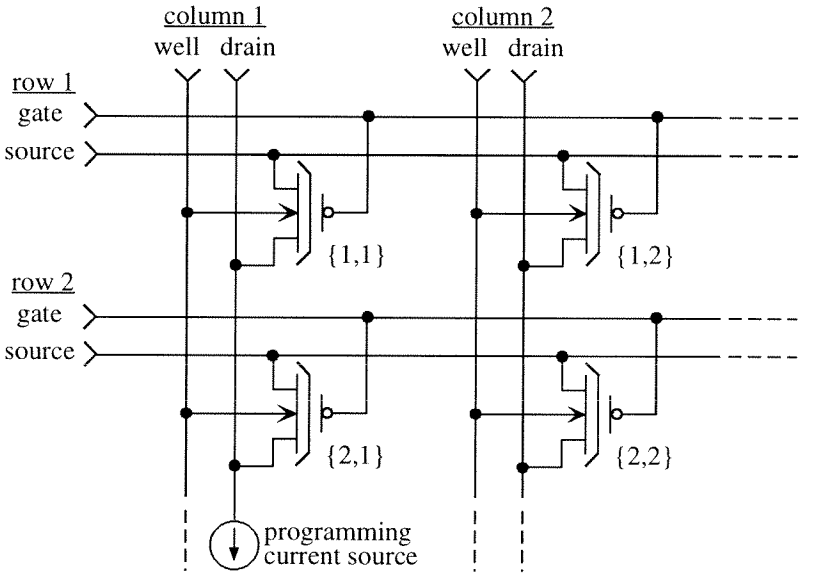


Figure 3.18 Self-convergent writes in a guarded-*p*FET array. The transistor symbol is a guarded-*p*FET synapse: The arrow on the well-contact wire denotes a tunneling junction, and the extensions to the floating-gate symbol denote guard rings. I use the circuit in (A) to write an analog drain-current value. As I describe in Section 3.2.5.1, I merely sink the programming current from the synapse transistor's drain. An intrinsic feedback loop, comprising the current source and the synapse transistor, adjusts the floating-gate charge to equalize the programming and *p*FET-drain currents. I fabricated the 2×2 array in (B) to measure the crosstalk between synapses during memory writes. The column synapses share a drain wire; consequently, when the current source pulls the {1,1} synapse's drain low to effect IIHEI, thermally generated carriers can inject onto the {2,1} synapse's floating gate. The write crosstalk, defined as the percentage change in the {2,1} synapse's drain-current value following a full-scale write of the {1,1} synapse, is about 0.025%

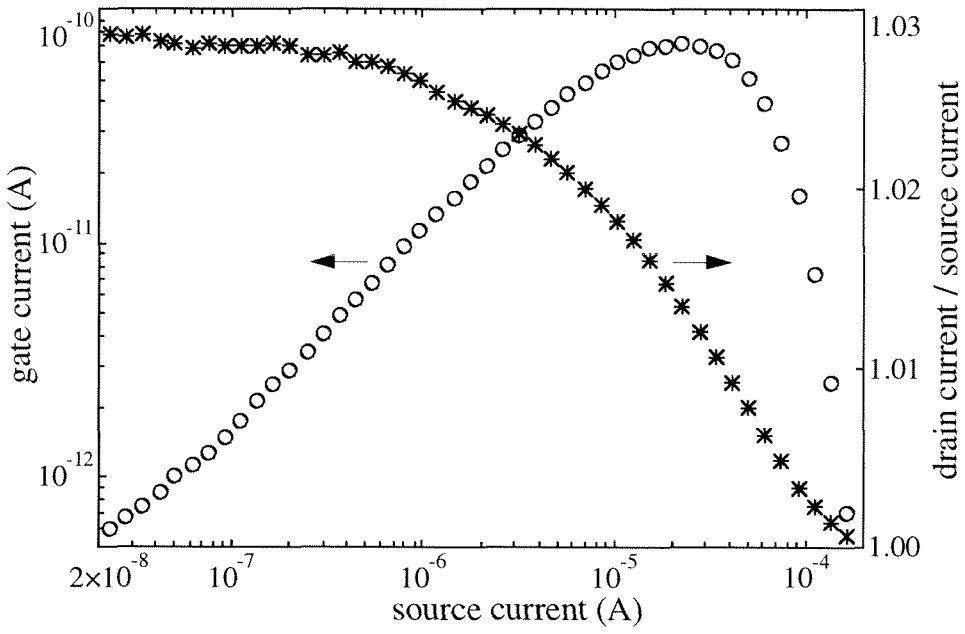


Figure 3.19 Maximizing a guarded-*p*FET synapse's write rate. For the synapse of Figure 3.11, I show the gate current and the impact-ionization efficiency, versus the source current, for a fixed drain-to-source voltage V_{ds}=12V. In an analog EEPROM, I avoid source currents smaller than about 20nA, because the IIHEI gate current, and therefore the write rate, are small. The *p*FET's transconductance changes rapidly near threshold; consequently, I further avoid source currents smaller than about 200nA, to minimize the read–write gain and nonlinearity errors described in Section 3.2.5.1 and shown in Figure 3.21. At high source currents, the potential at the drain end of the channel drops, thereby decreasing the drain-to-channel electric field and with it the impact-ionization probability. Above I_s=200μA, there are no impact-generated electrons, so there is no gate current. I employ drain-current values in the 200nA to 20μA range, thereby minimizing the read–write errors and maximizing the EEPROM write rate.

In an analog EEPROM, a primary concern is memory-write speed and accuracy. The silicon-MOS physics does not restrict synapse transistors to subthreshold source currents; in fact, all my synapse transistors exhibit large hot-electron gate currents for a wide range of above-threshold source currents. I use above-threshold source currents in the analog EEPROM, to speed the write process. In Figure 3.19, I show gate current versus source current for a guarded-*p*FET synapse with source currents ranging from 20nA to 200μA. These data show that I can maximize the memory-write speed by using source currents in the 2μA to 20μA range.

I wrote 64 logarithmically spaced drain-current values to a memory cell, using a 100msec write pulsewidth, and I measured the write error for each write. In Figure 3.20, I show the read–write transfer function, and the write error, versus the write current. I repeated the experiment of

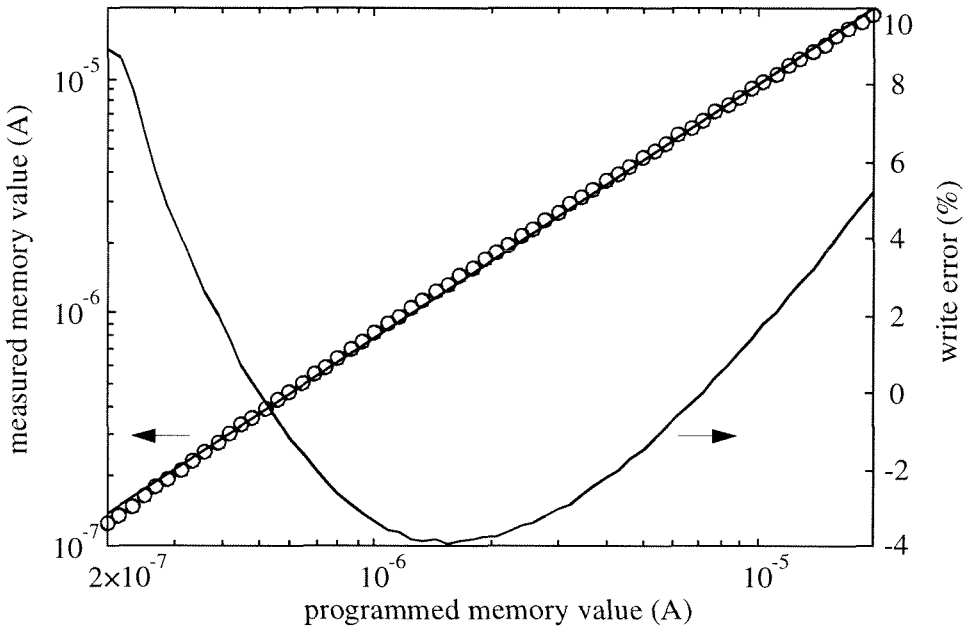


Figure 3.20 Read–write transfer function and write error, for a 100ms write pulsewidth. I wrote 64 logarithmically spaced drain-current values to the {1,1} synapse in Figure 3.18 (B). I chose log-scale values to illustrate the memory cell's dynamic range. I reset the drain current to 100nA prior to each write. I calculated a best-fit transfer-function line to the data, and plotted both the best-fit line and the fractional deviation between the fit line and the data. During cell erasure, if excessive tunneling occurred, the drain current became small; when I later wrote the cell, the gate current was small, and the write process was slow. I therefore initialize the cell after tunneling by (1) applying a 100nA programming current, (2) lowering the control-gate voltage until the drain current was equal to this programming current, and (3) using the self-convergent feedback mechanism, described in Section 3.2.5.1, to maintain this drain-current value as I ramped the control gate back up to its resting voltage. I can initialize an entire array by sinking 100nA from every column, and successively ramping the control-gate inputs for each row.

Figure 3.20, for write pulsewidths ranging from 68msec to 10sec; in Figure 3.21, I show the read–write gain and nonlinearity errors, versus the write pulsewidth. To prevent writing a cell during memory reads, I used lower drain voltages for reading than I did for writing. As a result of the synapse's parasitic floating-gate–to–drain overlap capacitance, this drain-voltage differential coupled to the floating gate, causing an offset between the write current and the read current. Because the *p*FET's transconductance is nonlinear, this offset varied with the memory-write value. Also, the shorter the programming pulsewidth, the further the drain voltage was from its settled value when I removed the programming current, and the larger the errors. To reduce these write errors in future arrays, I can sense the drain voltage during writing, and can always disable the

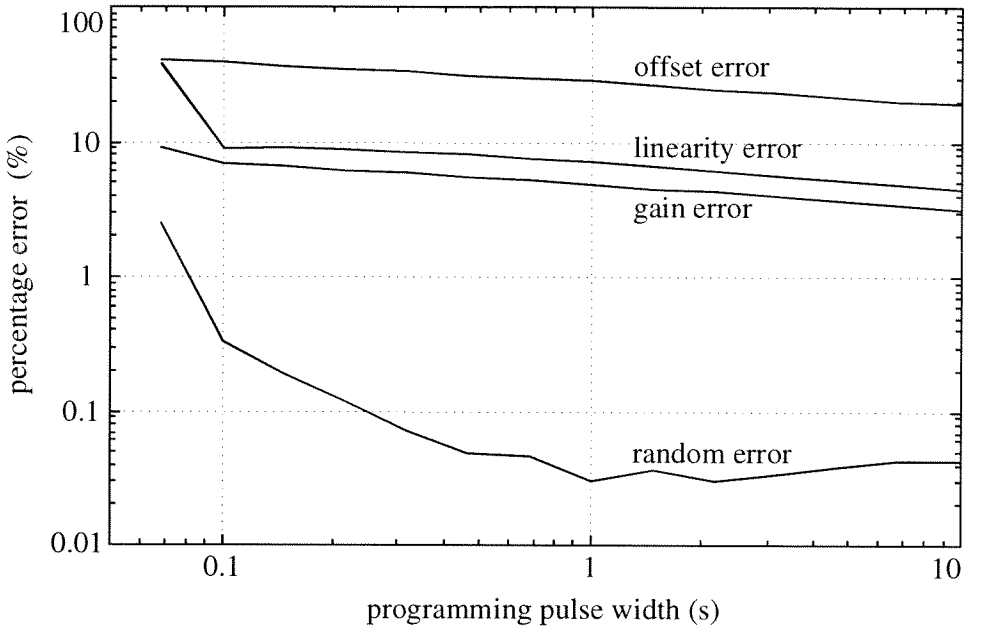


Figure 3.21 Memory-cell write errors versus write pulsewidth. I repeated the experiment of Figure 3.20, using write pulsewidths ranging from 68 msec to 10 sec. I plotted the offset error (the maximum deviation between any stored drain-current value and the respective programming current), the linearity error (the maximum deviation between any stored drain-current value and the best-fit transfer-function line), the gain error (the deviation of the best-fit transfer-function line from unity slope), and the random error (the RMS write error after removing the nonlinearity) versus the write pulsewidth. Because I employed oversized (1 pF) gate capacitors, and used an off-chip current source to write the memory, the settling times were long. The shorter the programming pulsewidth, the further the drain voltage was from its settled value when I removed the programming current, and the larger the errors.

write when the drain reaches a fixed voltage. This change will ensure consistent write errors that can be compensated by circuitry at the array boundary.

3.2.5.2 Erasing the Memory

A guarded-*p*FET EEPROM permits either flash or single-cell erasure, depending on the chip layout. If I fabricate an entire array within a single *n*-type well, I can flash erase the rows. I select a row for erasure by applying a high tunneling voltage to the *n*well, and a low voltage to the control-gate input of the selected row. If I instead fabricate each column of the array within its own *n*-type well, I can erase cells individually. I erase an individual cell by applying a high tunneling voltage to the column *n*well, and a low voltage to the row control-gate input. Regardless of the *n*well layout, the array permits single-cell writes.

3.2.5.3 Fabricating the EEPROM in Standard CMOS Processes

Guarding the source and drain junctions of *n*-type MOSFETs is impractical, because the floating-gate voltage would need to be near the tunneling voltage. Instead, EEPROM vendors employ special implants and processing steps to permit high-voltage tunneling in unguarded *n*-type MOSFETs. This requirement for specialized processing has prevented the wide use of floating-gate devices in conventional MOS design. However, guarding the source and drain junctions of *p*-type MOSFETs is trivial, as I have already shown, and *p*-type MOSFETs exhibit the further benefit of self-convergent memory writes. Guard rings and floating-gate transistors can be fabricated in any MOS process (although a double-poly process simplifies the layout). Consequently, my guarded-*p*FET synapse can be made a standard element in CMOS integrated circuit design, thereby allowing the integration of nonvolatile-analog or multilevel-digital storage in standard CMOS processes.

3.3 Further Development

My *p*FET synapses already possess those attributes that I believe are essential for building a silicon-learning system. In addition, the guarded-*p*FET synapse can be used as a nonvolatile analog-storage element in conventional CMOS design. However, further development, especially in the four areas that I have discussed for the *n*FET synapses (see Section 2.3), will improve my *p*FET synapses substantially. These key areas remain (1) reduced tunneling voltages, (2) reduced tunneling-junction leakage, (3) reduced overlap capacitances, and (4) smaller synapse size. Like for the *n*FET synapses, more modern processing will readily allow these improvements.

References

1 C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "Hole impact-ionization method of hot-electron injection and a four-terminal *p*FET semiconductor structure for long-term learning," provisional patent application submitted to the U.S. Patent Office on 29 April, 1996, and assigned serial no. 60/016,464; utility patent application submitted to the U.S. Patent Office on 22 April, 1997.

2 D. Frohman-Bentchkowsky, "Memory behavior in a floating-gate avalanche-injection MOS (FAMOS) structure," *Appl. Phys. Lett.,* vol. 18, no. 8, pp. 332–334, 1971.

3 A. S. Grove, *Physics and Technology of Semiconductor Devices*, New York: John Wiley & Sons, 1967.

4 K. Ohsaki, N. Asamoto, and S. Takagaki, "A single poly EEPROM cell structure for use in standard CMOS processes," *IEEE J. Solid-State Circuits,* vol. 29, no. 3, pp. 311–316, 1994.

5 C. Bleiker and H. Melchior, "A four-state EEPROM using floating-gate memory cells," *IEEE J. Solid-State Circuits,* vol. sc-22, no. 3, pp. 460–463, 1987.

6 H. V. Tran, T. Blyth, D. Sowards, L. Engh, B. S. Nataraj, T. Dunne, H. Wang, V. Sharin, T. Lam, H. Nazarian, and G. Hu, "A 2.5 V 256-level non-volatile analog storage device using EEPROM technology," in *Proc. 1996 IEEE Intl. Solid-State Circuits Conf.,* Dig. Tech. Papers, San Francisco, CA, pp. 270–271, 1996.

7 P. Hasler, B. A. Minch, C. Diorio, and C. Mead, "An autozeroing amplifier using *p*FET hot-electron injection," in *Proc. 1996 IEEE Intl. Symp. on Circuits and Systems*, Atlanta, vol. 3, pp. 325–328, 1996.

Chapter 4

A Floating-Gate MOS Learning Array with Locally Computed Weight Updates

Hebb's postulate—that synapse growth occurs as a result of coincident presynaptic and postsynaptic activity—can form the basis for local learning in an array of silicon synapse transistors. I show a block diagram of a candidate array in Figure 4.1. This array computes the inner product of an input vector **X** and the stored analog weight matrix. The computation and synapse-weight modification occur locally and in parallel: Column inputs (the presynaptic signals) that are coincident with row-learn signals (the postsynaptic activity) cause weight increases at selected synapses. To prevent unbounded weight values, I constrain the synapse weights using a fed-back row-error signal.

I have fabricated a 4×4 array of four-terminal *n*FET synapse transistors based on the architecture of Figure 4.1. I chose the input vector **X** and the row-learn vector **Y** to comprise 10μs digital pulses. To simplify the testing and subsequent analysis, I chose each row-learn signal Y_i to match one of the column inputs X_j identically; consequently, the learning is not Hebbian. In future testing, however, I can derive each row-learn signal Y_i from the ith row output, thereby implementing a true Hebbian learning rule. I chose a simple constraint to bound the weight values: The time-averaged sum of the synapse weights, in each row of the array, is held constant. This constraint forces row synapses to compete for floating-gate charge, stabilizing the learning.

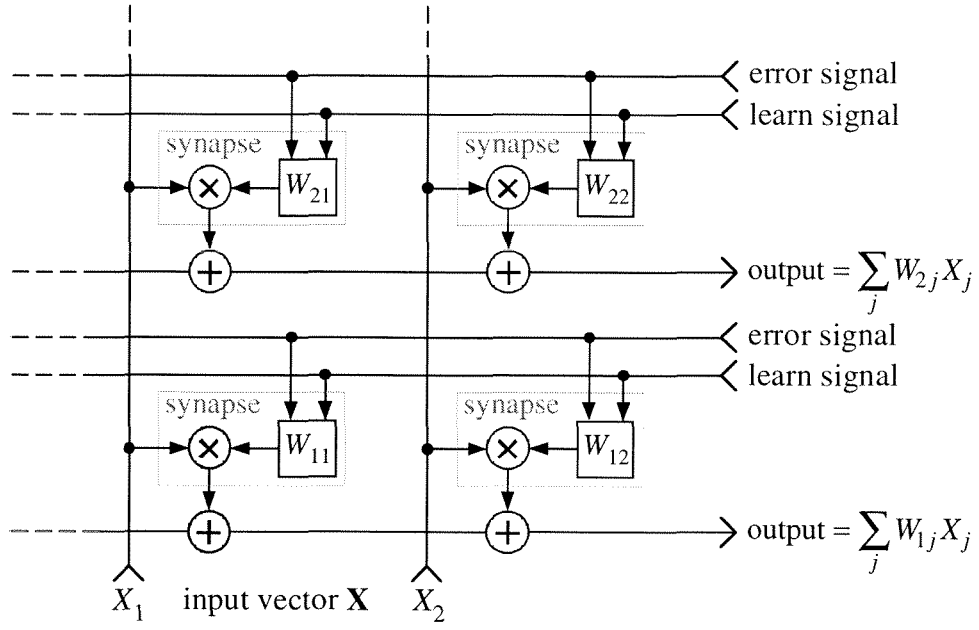


Figure 4.1 The learning-array block diagram. For clarity, I show only a 2×2 block of the 4×4 array. Each synapse multiplies its column input with its stored analog weight, and outputs a current to the row output wire; the row wire sums the synapse-output currents along the row. The stored weights are nonvolatile; column inputs that are coincident with row-learn signals cause weight increases at selected synapses. The error signal constrains the time-averaged sum of the row-synapse weights to be a constant, bounding the row weights by forcing the synapses to compete for weight value.

4.1 The Learning Array

In Figure 4.2, I show one row of the learning array, comprising a synapse transistor at each array node and a normalization circuit at the row boundary. The column inputs X_i and the row-learn signals Y_j are 10μs digital pulses. Each synapse multiplies its binary-valued input X_i with its stored weight W_{ij}, and outputs a source current I_{sij} whose magnitude is given by Eqn. (1.2). The total row current I_{out} is the sum of the source currents from all the synapses in the row. Synapses ordinarily are on; low–true gate inputs X_i turn off selected synapses, decreasing the current I_{out} transiently. This decrease in I_{out}, in response to an input vector **X**, is the row computation.

Synapse-weight increases occurs only when both the row and column inputs, Y_j and X_i, are true. To see why, I first consider the case when the row learn signal Y_j is false (V_{tun} is low). Because $V_{ox} \equiv V_{tun} - V_{fg}$, when V_{tun} is low, V_{ox} is small for every synapse in the row. When V_{ox} is small, the tunneling currents are small, and there is no weight increase at any row synapse.

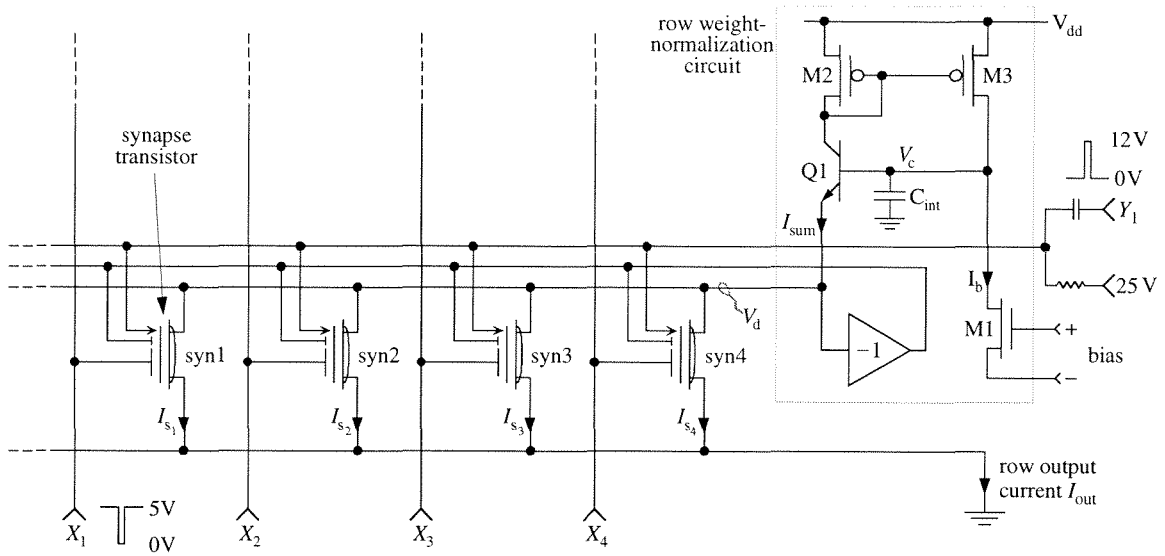


Figure 4.2 One row of the learning array. The column input vector **X** comprises low–true, 5V, 10μs digital pulses; the row input vector **Y** comprises high–true, 12V, 10μs digital pulses. Because the 2μm CMOS process that I use has 400Å gate oxides, the tunneling voltages are high; to cause measurable tunneling, I superimpose the row inputs onto a 25V DC bias. The voltage coupling between a synapse's control and floating gates is about 0.8. Consequently, a 5V (low–true) input on column wire X_1 causes a 4V decrease in syn1's floating-gate voltage; this decrease, in turn, causes a 4V increase in syn1's tunneling-oxide voltage. A column input X_1 that is coincident with a row learn pulse Y_1 causes a 16V increase in the tunneling-oxide voltage at syn1, but causes only a 12V increase at the other synapses. Because electron tunneling increases exponentially with the tunneling-oxide voltage (see Figure 2.2), syn1's floating gate receives about 100 times more charge than do the other synapses' floating gates; because W increases exponentially with the floating-gate charge (see Eqn. (1.2)), syn1's weight increases much more than do the other synapses' weights. The weight increase causes I_{sum} to rise, which, in turn, causes the normalization circuit to raise V_d. Because the CHEI efficiency increases with V_{dc} (see Figure 2.4), a higher V_d causes CHEI in all the synapses, decreasing all the weights. The array eventually settles back to equilibrium, with I_{sum} equal to I_b, but syn1 now takes a larger share of the total row current, and the other synapses each take a smaller share. The inverting amplifier in the weight-normalization circuit enhances loop stability, for reasons that I discuss in Section 4.3.3.

Now I consider the case when Y_j is true (V_{tun} is high). V_{ox} increases with as V_{fg} decreases, and V_{fg} follows X_i. If a low–true column input X_i is true, then V_{fg} is low; V_{ox} is large, and electron tunneling causes a weight increase at the selected synapse. If, on the other hand, a low–true column input X_i is false, then V_{fg} is high; V_{ox} is too small to cause appreciable tunneling, and there is little change in the synapse's weight.

Tunneling increases the weight value of a row–column selected synapse. Because this weight update is single quadrant, tunneling allows unbounded weight increases. To constrain the array-weight values, I renormalize the weights in each row of the array. My array affords unsupervised learning [1], with the following constraint: The sum of the row-synapse weights, averaged over time, is a constant. The array error metric is a weight normalization; I use CHEI feedback along each row of the array to enforce the constraint.

4.2 Weight Normalization

The weight-normalization circuit (see Figure 4.2) compares I_{sum}, the sum of the synapse drain currents in the row, with I_b, the bias current in transistor M1; if $I_{sum}>I_b$, then the circuit uses CHEI to renormalize the weights. To explain the renormalization, I begin by defining row equilibrium: A row is in equilibrium when $I_{sum}=I_b$. In equilibrium, the drain voltage V_d typically causes little or no CHEI in the row synapses.

The normalization circuit constrains I_{sum} as follows: Assume that the row initially is in equilibrium, and that tunneling then increases the weight values of selected synapses, increasing I_{sum}. The excess drain current ($I_{sum}-I_b$) is mirrored by M2 and M3 into capacitor C_{int}, causing V_c to rise; Q1 forces V_d to follow V_c. When V_d rises, all the row synapses undergo CHEI, decreasing all the weights, causing I_{sum} to fall. As I_{sum} falls, V_d also falls, and the row returns to equilibrium. The drain-current constraint requires that, over time, $I_{sum}=I_b$. The normalization circuit creates a negative resistance at the synapses' common drain node, causing V_d to rise when I_{sum} increases.

4.2.1 The Drain-Current Constraint Renormalizes the Weights

I now show how the drain-current constraint renormalizes the row-synapse weights. I begin with the constraint

$$\sum_i I_{s_i} \approx \sum_i I_{d_i} \equiv I_{sum} = I_b \tag{4.1}$$

In Section 4.4, I show that the renormalization time constant τ_a exceeds 10s; this value is 10^6 times longer than the 10μs input pulses X_i (where $V_{in}=X_i$). Consequently, for renormalization, I replace V_{in} in Eqn. (1.2) with its temporal average $\overline{V}_{in}$, and I assume that $\overline{V}_{in}$ both is time invariant and has the same value for all the row synapses. I then substitute Eqn. (1.2) into Eqn. (4.1):

$$\sum_i W_i I_o e^{\frac{\kappa' \overline{V}_{in}}{U_t}} = I_o e^{\frac{\kappa' \overline{V}_{in}}{U_t}} \sum_i W_i = I_b \tag{4.2}$$

$$\Rightarrow \quad \sum_i W_i = \frac{I_b}{I_o} e^{\frac{-\kappa' \overline{V}_{in}}{U_t}} \equiv W_{sum} = \text{constant} \tag{4.3}$$

The drain-current and weight-value constraints are equivalent; consequently, row feedback renormalizes the synapse weights.

Renormalization forces the row synapses to compete for floating-gate charge; when one synapse's weight value increases, the sum of the weight values of its row neighbors must decrease by the same amount. However, when a selected synapse tunnels, increasing its weight, renormalization forces *all* the row synapses to undergo CHEI, decreasing *all* the row-synapse weights. The selected synapse undergoes both tunneling and CHEI; because the exponent in the CHEI weight-decrement rule is larger than that in the tunneling weight-increment rule (see Eqns. (2.17) and (2.7), respectively), renormalization constrains a synapse's weight-update rate, in addition to its weight value.

4.2.2 The Array Learning Rule

Tunneling and CHEI effectively redistribute a fixed quantity of floating-gate charge among the row-synapse transistors. I now derive the array learning rule, for coincident (x,y) pulse inputs to synapse j. I consider the row-synapse weights at discrete time intervals $t \equiv n\mathrm{T}$, where n is the step number and T is the timestep, and derive the row-learning rule for a single coincident (x,y) input to a single row synapse. I begin with the equilibrium condition for the row-weight normalization:

$$\sum_i W_i(n) = W_{sum} \tag{4.4}$$

I assume that the normalization time constant τ_a is fixed, for the following reason. Coincident (x,y) input pulses cause a weight increase at a synapse; the normalization circuit responds by establishing a drain voltage V_d for which the total weight decay, summed over all the row synapses, balances the weight increase at the single synapse. If I assume that the mean density of the coincident input pulses is time invariant, then V_d's mean value, $\overline{V}_d$, is constant, and therefore the low-frequency loop time constant τ_a also is constant.

I further assume that $\tau_a \ll \mathrm{T}$. The synapse-weight values can violate Eqn. (4.4) for times $t \ll \tau_a \ll \mathrm{T}$, but I require that they satisfy Eqn. (4.4) at my measurement time intervals $t = n\mathrm{T}$. I permit array inputs at times $(t+\delta t) \equiv (n+\delta)\mathrm{T}$, immediately after I measure the synapse-weight values at $t = n\mathrm{T}$. The array inputs comprise a pulsed column vector $\mathbf{X}(n+\delta)$, where $X_i \in [0,1] \equiv [5\mathrm{V}, 0\mathrm{V}]$, and a pulsed row vector $\mathbf{Y}(n+\delta)$, where $Y_j \in [0,1] \equiv [0\mathrm{V}, 12\mathrm{V}]$. Without loss of generality, I assume that, at time $t = n\mathrm{T}$, the circuit is in equilibrium, and that, at time

$(t+\delta t+t_{pw})\equiv(n+\delta_{pw})T$, coincident row and column inputs, of duration t_{pw}, have caused synapse j's weight to increase:

$$W_j(n+\delta_{pw}) \approx W_j(n)+\frac{\partial W_j(n)}{\partial t}t_{pw} \tag{4.5}$$

$$\approx W_j(n)+\frac{t_{pw}}{\tau_{tun}}W_j(n)^{(1-\sigma)} \tag{4.6}$$

where in Eqn. (4.5) I have made the first-order approximation that $\partial W/\partial t$ is constant over t_{pw}, and in Eqn. (4.6) I have substituted for $\partial W/\partial t$ using Eqn. (2.7). Because $t_{pw}<<\tau_a$, at time $(n+\delta_{pw})$ the circuit no longer is in equilibrium,

$$\sum_i W_i(n+\delta_{pw}) > W_{sum} \tag{4.7}$$

and the synapse weights inject down to reestablish equilibrium.

I wish to find the synapse weights at $(n+1)$, when the row again satisfies Eqn. (4.4). Using Eqns. (2.17) and (4.6), I write weight-decrement expressions for the row synapses,

$$\Delta W_{i,i\neq j}(n+1) = -\frac{T}{\tau_{inj\_4n}}W_{i,i\neq j}(n)^{(2-\varepsilon_{4n})} \tag{4.8}$$

$$\Delta W_j(n+1) \approx -\frac{T}{\tau_{inj\_4n}}\left(W_j(n)+\frac{t_{pw}}{\tau_{tun}}W_j(n)^{(1-\sigma)}\right)^{(2-\varepsilon_{4n})} \tag{4.9}$$

where, because the row drain voltage V_d settles during renormalization, τ_{inj_4n} may vary over T (recall that $T>>\tau_a>>t_{pw}$). For reasonable values of V_{tun} and t_{pw}, the weight increment from a single coincident (x,y) input is small; consequently, I can simplify Eqn. (4.9) using $(1+x)^n\approx 1+nx$,

$$\Delta W_j(n+1) \approx -\frac{T}{\tau_{inj\_4n}}W_j(n)^{(2-\varepsilon_{4n})}\left(1+(2-\varepsilon_{4n})\frac{t_{pw}}{\tau_{tun}}W_j(n)^{-\sigma}\right) \tag{4.10}$$

Because τ_{inj_4n} varies over T, I now re-express T/τ_{inj_4n} in terms of quantities that I know at n. I equate the weight increment at synapse j (see Eqn. (4.6)) to the sum of the weight decrements at synapses $i, i\neq j$ (Eqn. (4.8)) and j (Eqn. (4.10)):

$$\frac{t_{pw}}{\tau_{tun}}W_j(n)^{(1-\sigma)} = \frac{T}{\tau_{inj\_4n}}\sum_{i,i\neq j}W_i(n)^{(2-\varepsilon_{4n})} + \frac{T}{\tau_{inj\_4n}}W_j(n)^{(2-\varepsilon_{4n})}\left(1+(2-\varepsilon_{4n})\frac{t_{pw}}{\tau_{tun}}W_j(n)^{-\sigma}\right) \tag{4.11}$$

Then, I solve for T/τ_{inj_4n}:

$$\frac{T}{\tau_{inj\_4n}} = \frac{\frac{t_{pw}}{\tau_{tun}}W_j(n)^{(1-\sigma)}}{(2-\varepsilon_{4n})\frac{t_{pw}}{\tau_{tun}}W_j(n)^{(2-\varepsilon_{4n}-\sigma)}+\sum_i W_i(n)^{(2-\varepsilon_{4n})}} \tag{4.12}$$

I define $f_{learn}\equiv T/\tau_{inj\_4n}$, substitute f_{learn} into Eqn. (4.8), and use Eqn. (4.4) to solve for the row-learning rule:

$$W_{i,\,i\neq j}(n+1) = W_{i,\,i\neq j}(n) - f_{\text{learn}}\, W_i(n)^{(2-\varepsilon_{4n})} \tag{4.13}$$

$$W_j(n+1) = W_j(n) + f_{\text{learn}} \sum_{i,i\neq j} W_i(n)^{(2-\varepsilon_{4n})} \tag{4.14}$$

Eqns. (4.13) and (4.14) describe the row weight-update rule for a single coincident (x,y) pulse input to synapse j. In Figure 4.3 and Figure 4.4, I show unsupervised learning in one row of my 4×4 array; these data highlight both the synapse weight and the update-rate constraints. I fit these data by applying Eqns. (4.13) and (4.14), recursively; the only inputs to the fit equations are the synapse weights at $n=0$ and the fit constants τ_{tun}, t_{pw}, σ, and ε_{4n}.

4.3 Normalization-Circuit Stability

The normalization circuit creates a negative resistance at the synapses' common drain node: When I_{sum} increases, V_{d} rises. The loop output is V_{d}, and the loop feedback comprises CHEI oxide currents: When V_{d} rises, CHEI decreases the synapse weights, causing I_{sum} to fall. Because the CHEI oxide currents increase exponentially with V_{d}, the loop dynamics are highly nonlinear. I therefore describe qualitative, rather than quantitative, loop-stability criteria.

The normalization circuit employs positive feedback; to ensure stability, I must make the loop gain less than unity for all frequencies. This requirement implies that the small-signal impedance z_{d}, looking into the synapse drain terminals, must be greater than the total impedance z_{c}, at capacitor C_{int}. To see why, I assume instead that $z_{\text{c}}>z_{\text{d}}$. A rising V_{d} induces a small-signal current $i_{\text{sum}}=v_{\text{d}}/z_{\text{d}}$; i_{sum} is mirrored by M2 and M3 into C_{int}, causing V_{c} to rise by an amount $v_{\text{c}}=i_{\text{sum}}z_{\text{c}}=(z_{\text{c}}/z_{\text{d}})v_{\text{d}}$. If $z_{\text{c}}>z_{\text{d}}$, then $v_{\text{c}}>v_{\text{d}}$; because V_{d} follows V_{c}, i_{sum} will increase rapidly, causing V_{c} to rise toward V_{dd}.

The impedance z_{d} is limited by interconnect capacitances, and by synapse-transistor channel-length modulation, floating-gate–to–drain overlap capacitance, and drain-current impact ionization. I consider each of these limitations in turn.

4.3.1 Interconnect Capacitance

Interconnect capacitance at the synapses' common drain node causes z_{d} to decrease with frequency. I chose C_{int} to be much larger than this parasitic capacitance, so the reactive impedance ratio, $z_{\text{c}}/z_{\text{d}}$, favors loop stability for all frequencies.

4.3.2 Channel-Length Modulation

Channel-length modulation reduces a synapse transistor's drain impedance, limiting z_{d}. Fortunately, the four-terminal nFET synapse's Early voltage exceeds 100V, as a result of both the

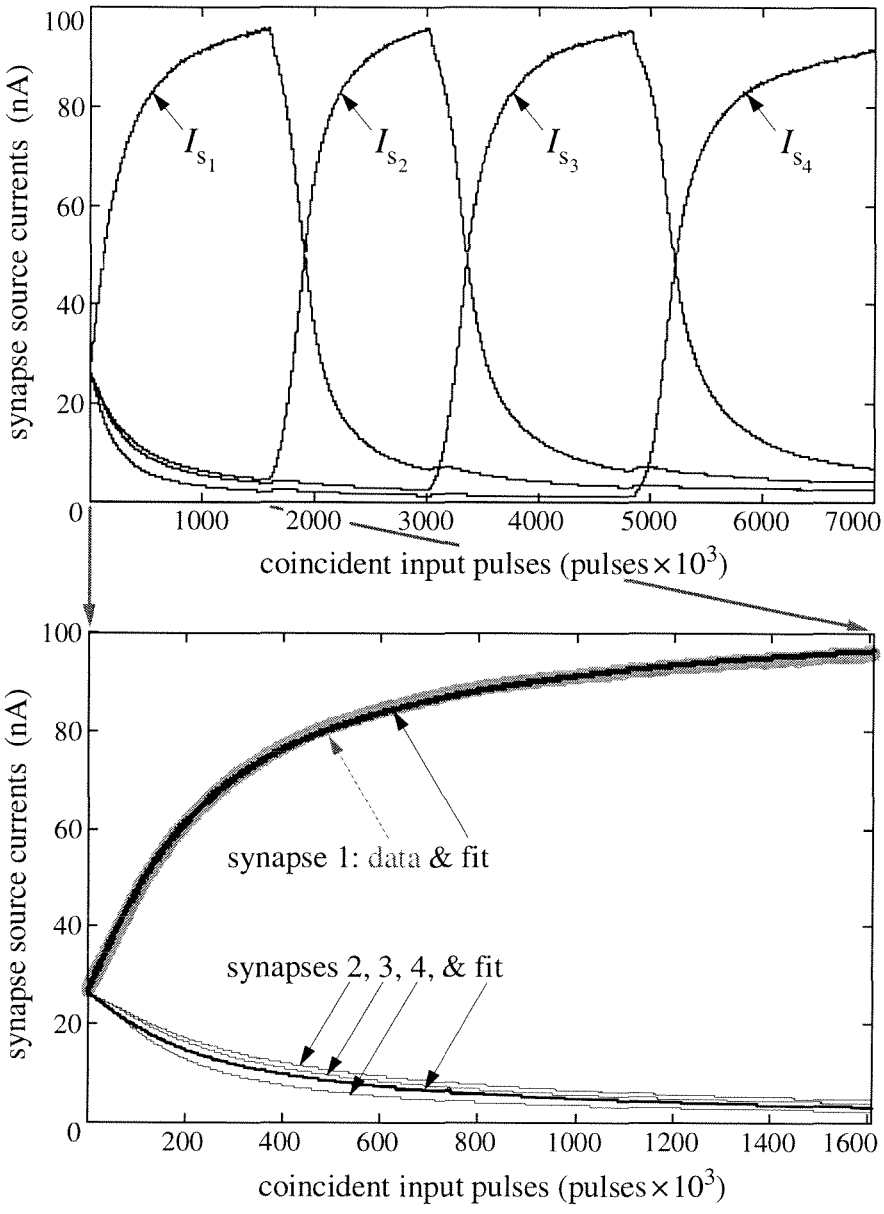


Figure 4.3 Array learning behavior, with fits. I initialized all four synapses to the same source-current value prior to starting the experiment. I first applied a train of coincident (x,y) 10μs pulses to synapse 1, causing synapse 1's weight value and source current to increase. Renormalization caused the weight values and source currents of the other synapses to decrease. Once synapse 1 had acquired 90% of the total row current I_{sum}, I removed the pulse-train stimulus and applied it instead to synapse 2, and then, in turn, to synapses 3 and 4. I measured the synapse source currents after every 10^3 input pulses. In the lower half of the figure, I highlight the first 1600 data points, and fit these data by applying Eqns. (4.13) and (4.14), recursively. The inputs to the fit equations are the initial synapse source-current values (at n=0); the pulsewidth t_{pw}=10μs; and the empirical constants τ_{tun}=10ms, σ=0.14, and ε_{4n}=0.21. These data show that I can address individual synapses with good selectivity, and can achieve wide separation in the weight values of selected versus deselected synapses.

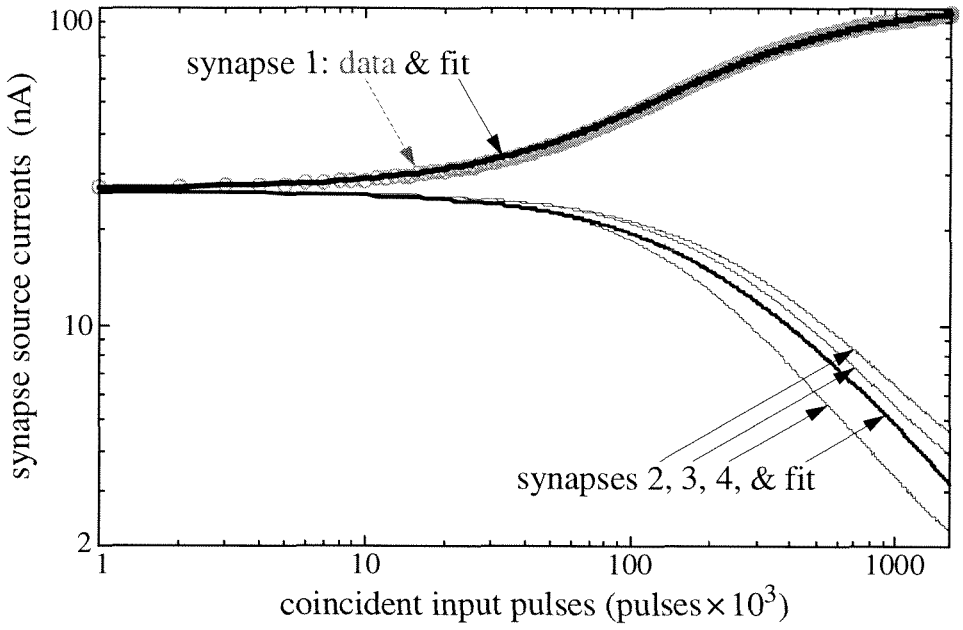


Figure 4.4 Logarithmic plot of the array learning behavior, with fits. I replotted the lower half of Figure 4.3, this time on a logarithmic, rather than on a linear, scale. This plot highlights both the synapse weight and growth-rate constraints, and shows that the weight values of deselected synapses do not saturate, but instead follow a power-law decay as predicted by Eqns. (2.17) and (4.13).

10μm channel length and the *p*-type channel implant; consequently, the channel-length modulation is small.

4.3.3 Floating-Gate–to–Drain Overlap Capacitance

The drain voltage V_d couples to a synapse transistor's floating gate, by means of the floating-gate–to–drain overlap capacitance C_{dg}. The coupling coefficient is C_{dg}/C_T, where C_T is the total floating-gate capacitance. Because I_s increases exponentially with V_{fg}, C_{dg} causes I_{sum} to increase exponentially with V_d, limiting z_d. To minimize this effect, I use a large interpoly capacitor ($C_T = 1\,pF$); I also apply inverting feedback from V_d to the floating gate, increasing z_d (see Figure 4.2). I use an off-chip amplifier to generate this inverting feedback; in future arrays, I will use instead an on-chip adaptive floating-gate amplifier [2].

4.3.4 Drain-Current Impact Ionization

Channel electrons that possess sufficient energy for CHEI also possess sufficient energy for impact ionization (see Section 2.1.5); consequently, a drain-to-channel electric field that induces weight renormalization also creates additional electron–hole pairs, causing I_d to increase expo-

nentially with V_{dc}. As a result, I_{sum} increases exponentially with V_d, limiting z_d. If V_d becomes greater than about 4V, the rate of the ionization-induced drain-current increase causes loop instability, and V_d rises rapidly. As V_d rises, CHEI decreases all the synapse-transistor weights; as V_d saturates near V_{dd}, CHEI causes I_{sum} to fall below I_b, causing V_d to fall, and the loop to return to a stable operating regime. Loop instability causes V_d to undergo a single brief (~10μs) voltage spike, and reduces all the synapse weights substantially. Fortunately, because the four-terminal *n*FET-synapse CHEI efficiency is high, weight renormalization rarely causes V_d to exceed 3.5V; consequently, the loop is stable.

4.4 Normalization-Circuit Response

In Figure 4.5, I show the normalization circuit's impedance magnitude versus frequency; in Figure 4.6, I show the circuit's impulse response. The impedance-versus-frequency plot shows that the low-frequency time constant τ_a (the adaptation time constant) typically exceeds 10s. The impulse-response plot shows that, for short timescales, the total drain current I_{sum} can exceed I_b, violating the normalization constraint. For long timescales, $I_{sum}=I_b$.

The parasitic coupling between a synapse's tunneling junction and its floating gate is about 5fF. With $C_T=1$pF, a 12V row-learn pulse Y_j transiently increases the floating-gate voltage of every row synapse by about 60mV. This coupling does not affect the row computation significantly, for two reasons. First, 5V low–true column inputs X_i always turn off selected synapses, regardless of Y_j. Second, because a row-learn pulse Y_j increases the floating-gate voltage of every deselected synapse by a fixed 60mV, I can calculate the corresponding source-current increase using Eqn. (1.1), and I can adjust I_{out} accordingly.

4.5 Further Development

I have demonstrated a silicon integrated circuit in which computation and weight modification occur locally and in parallel. The array achieves my goals of fast, single-transistor analog computation and of slow, locally computed weight adaptation: The inner product computes in 10μs, whereas the weight normalization takes minutes to hours.

I claim that, by deriving the row-learn signal from the row output I_{out}, I will, in future experiments, be able to demonstrate a true Hebbian learning rule in a silicon integrated circuit. Furthermore, although my array affords unsupervised learning, it uses a feedback error signal to constrain the weight values. Feedback error signals typically are used in supervised neural networks, to adjust the array weights according to the network learning rule. In future floating-gate

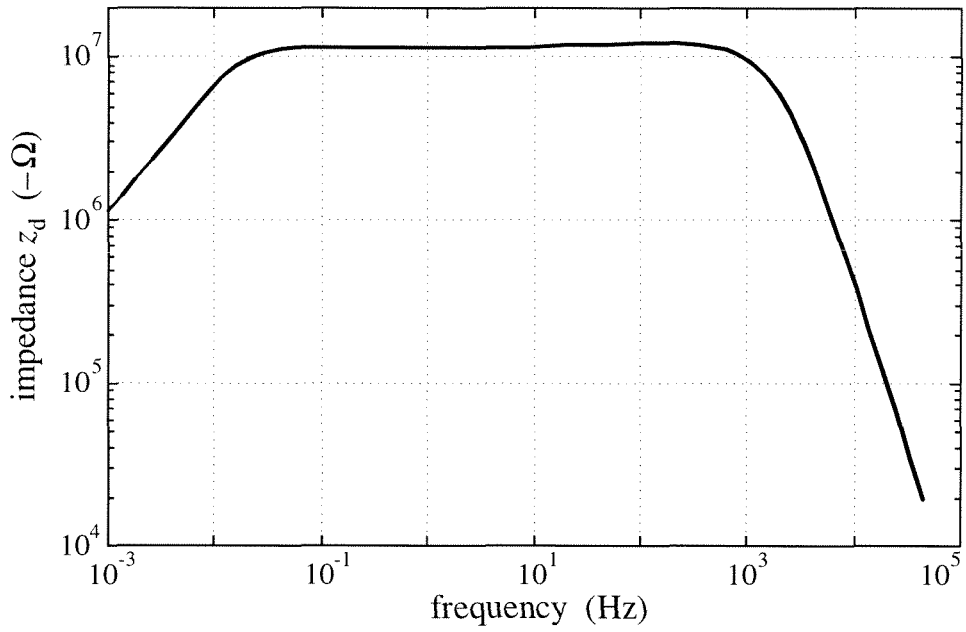


Figure 4.5 Normalization-circuit impedance magnitude versus frequency. I applied a small-signal sinusoidal current i_{in} to the synapses' row-drain node (see Figure 4.2), measured the resulting small-signal voltage v_d, and plotted $z_d = v_d / i_{in}$. Because the loop feedback comprises CHEI oxide currents, which increase exponentially with V_d, the low-frequency corner increases with V_d. To hold this corner at a single frequency, I applied a constant $V_{tun} = 37\,V$ to all the row-synapse transistors, causing continuous tunneling. The normalization loop re-established equilibrium by setting $V_d \sim 3.3\,V$, inducing continuous CHEI to compensate the continuous tunneling. For these (artificial) operating conditions, the low-frequency corner comprised a single pole at about 0.03 Hz. The high-frequency rolloff comprised two poles: The first was the normalization-loop response, set by C_{int}; the second was a consequence of interconnect capacitance at the synapses' common drain node, attenuating the injected signal i_{in}.

arrays, rather than using unsupervised learning, I intend to use CHEI to adjust the synapse weights in a supervised fashion, using either pulsed, or continuously valued analog [2], inputs and row-error signals.

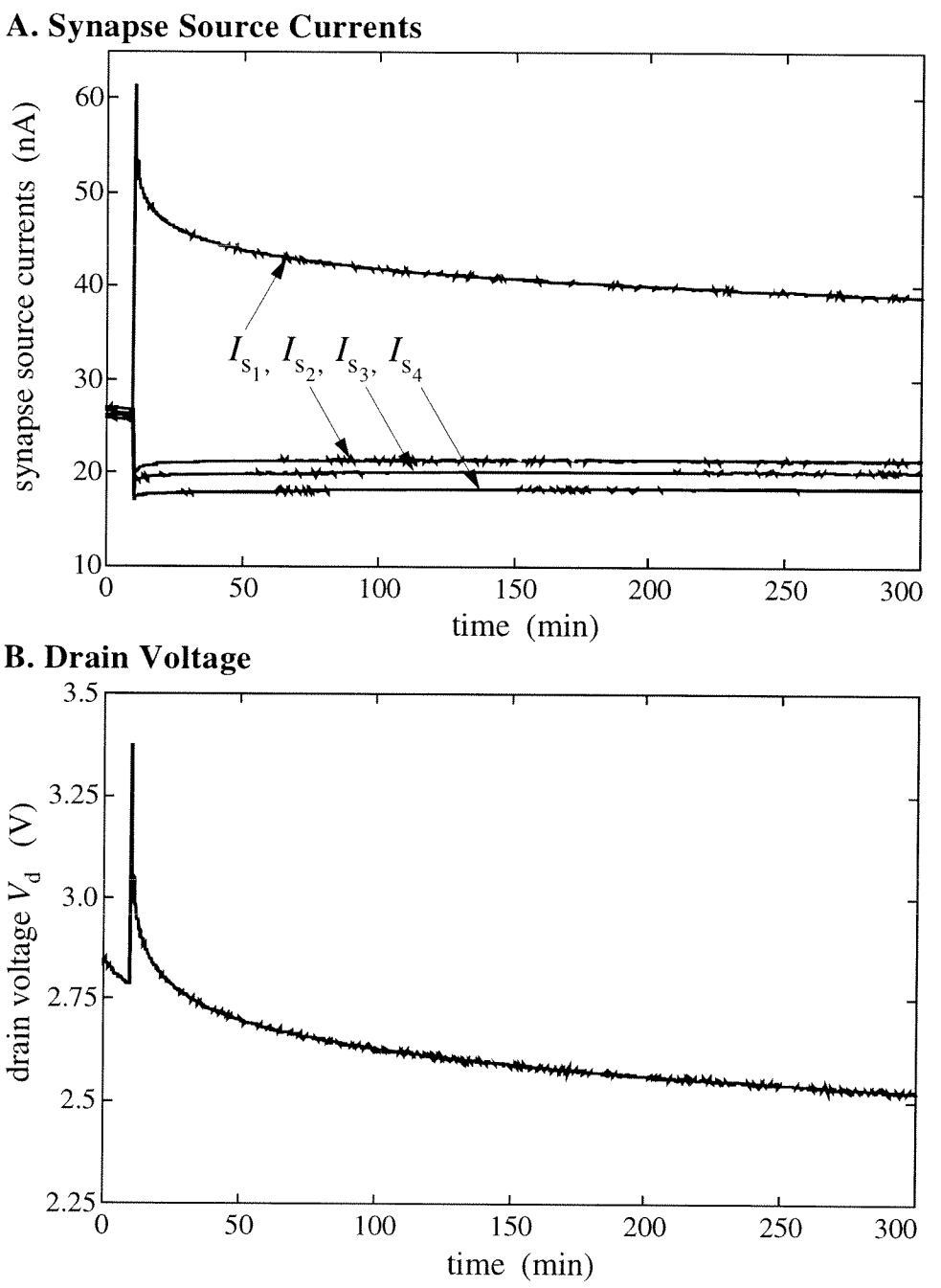


Figure 4.6 Normalization-circuit impulse response. At time t=10min, I applied 2×10^5 coincident (x,y) 10μs pulses, over a 10s period, to synapse 1. I plotted (A) the synapse source currents, and (B) the drain voltage V_d, for a period of about 5 hours following the stimulus. During CHEI, $\partial W/\partial t$ varies roughly as W^2; consequently, the source-current settling approximates a $1/t$ characteristic. (Note: because $\partial W/\partial t$ also varies with V_d, the settling differs somewhat from $1/t$.) After 2 weeks, V_d was about 2.4V. At time t=0, V_d initially was decaying, because I had just finished resetting the synapse source currents to identical values.

References

1 J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation,* Reading, MA: Addison-Wesley, 1994.

2 P. Hasler, *Foundations of Learning in Analog VLSI,* Ph.D. Thesis, Department of Computation and Neural Systems, California Institute of Technology, Pasadena, CA, 1997.

Chapter 5

Future Directions

5.1 Silicon Synapse Transistors

I have demonstrated single-transistor silicon synapses that store a nonvolatile analog weight, multiply an applied input by the stored weight, and allow locally computed weight updates. I have also demonstrated local, autonomous learning in a synaptic array. My colleagues and I have developed other synapse-transistor applications, not described here, including an analog memory cell that employs the four-terminal *n*FET synapse [1], and an autozeroing amplifier that employs the four-terminal *p*FET synapse [2].

I can build silicon neural networks using my synapse transistors. Unfortunately, despite my assertion that the inspiration for the neural-networks field derives from neuroscience, at present the field is better described by the term *network computation* than it is by *neural computation*. The dominant neural-network learning algorithm is back-propagation of errors, but there is scant neurophysiological evidence for back-propagating errors in neuronal memory formation [3], and many of the alternative neural-network learning algorithms are equally implausible biologically. Consequently, the neural-networks field emphasizes weights, rather than synapses, because the interconnections between node elements comprise primarily weight values. Although my silicon devices store a weight, I intentionally use the term synapse to describe them, because I have endeavored to embed in them some of the attributes of real neural synapses. I hope that, because of my work, the dividing line between network computation and neural computation, and between weights and neural synapses, will become blurred.

My goal, espoused in Section 1.3, is to build silicon circuits that exhibit behavior analogous to that of nervous tissue. I believe that my synapse transistors afford an essential building block toward achieving this goal. In Figure 5.1, I show post-tetanic potentiation (PTP) and long-term

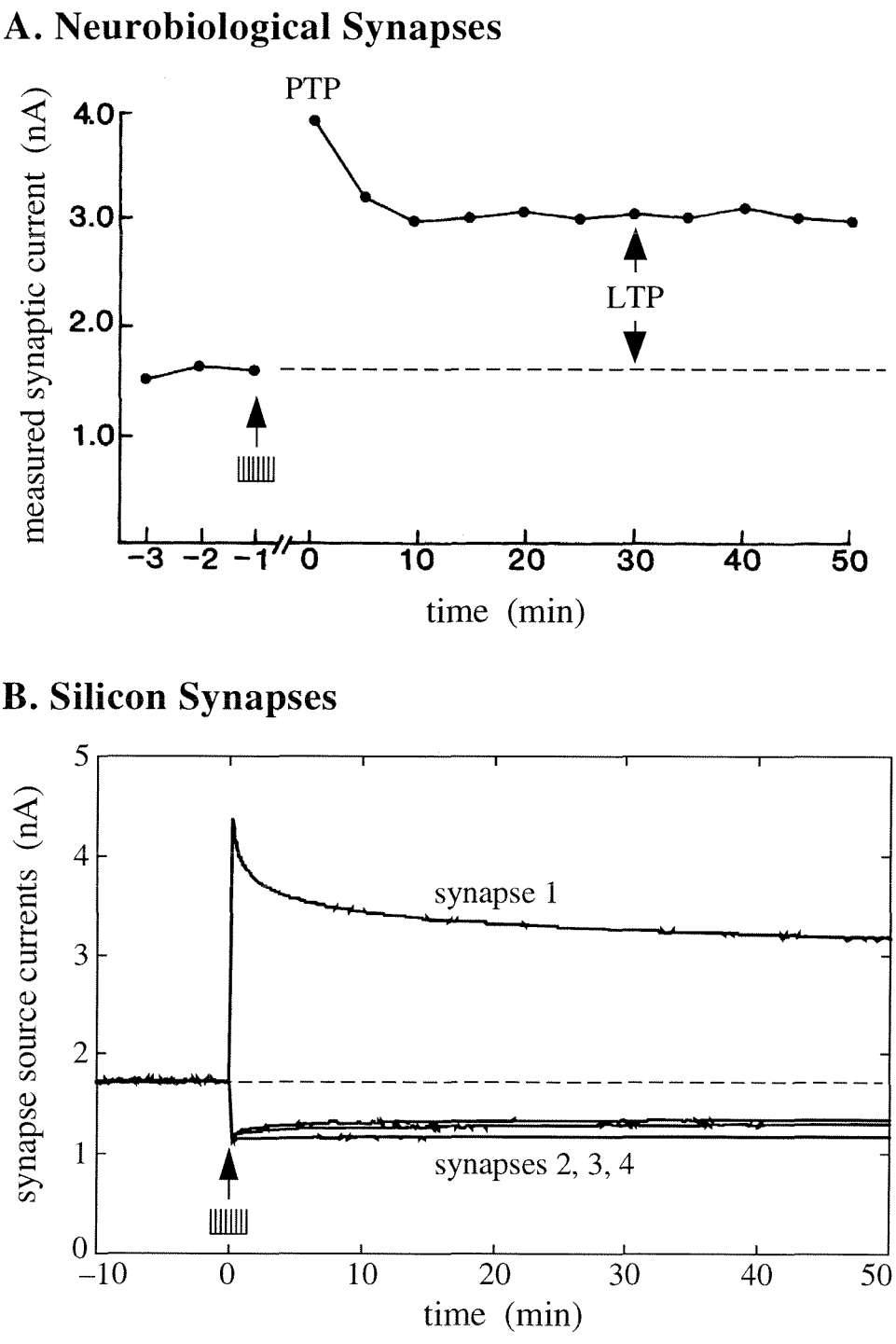


Figure 5.1 Post-tetanic potentiation (PTP) and long-term potentiation (LTP) in neurobiological and silicon synapses. (A) Measured mossy-fiber excitatory postsynaptic currents, before and after the induction of LTP, at a synaptic input to a CA3 pyramidal neuron of the disinhibited rat hippocampal slice. The data comprise average current-clamp recordings both before and after the application of a tetanic stimulus at the time indicated. *Source:* Adapted from G. Barrionuevo, S. R. Kelso, D. Johnston, and T. H. Brown, "Conductance

mechanism responsible for long-term potentiation in monosynaptic and isolated excitatory synaptic inputs to hippocampus," *J. of Neurophysiology,* vol. 55, no. 3, pp. 540–550, 1986, © American Physiological Society, used with permission. (B) Measured silicon-synapse source currents, before and after I applied a tetanic stimulus to synapse 1 in the learning array of Figure 4.2. The stimulus comprised 2×10^5 coincident (x,y) 10μs pulses over a 10s period. Synapse 1 exhibited behavior similar to that demonstrated by the neural synapse, with similar output-current values and a similar adaptation timescale. The depressive behavior of synapses 2 through 4 derives from the array learning rule (see Section 4.2.2); in the absence of equivalent neurobiological data, this depressive response is plausible but unsupported.

potentiation (LTP) in neurobiological and silicon synapses. These data show that silicon synapses can exhibit behavior similar to that demonstrated by neural synapses, with similar output-current values and a similar adaptation timescale. Although synapse transistors possess only a subset of the attributes of neural synapses, they can mimic some of the known behavior of neural synapses—most notably, the long-term autonomous learning.

Synapse transistors will make possible silicon learning systems modeled after neurobiology. In Figure 5.2, I show spike train recordings from neurobiological and silicon circuits. The neurobiological data were recorded from the somatosensory cortex of an anaesthetized cat. The silicon data were recorded from a simple oscillator that, although not neurally inspired, employs synaptic devices and positive feedback just like neurobiology, and exhibits behavior strikingly similar to the neurobiology. I do not mean to imply that the silicon circuit actually models the neurobiology—the action-potential train in part B of Figure 5.2 has meaning to the cat, whereas the oscillator output in part C is merely an interesting waveform. However, when I used neurally inspired components (synapse transistors) and neurally inspired circuit connections (positive feedback) in a silicon system, I obtained behavior that looked like neurobiology. These data are consistent with my belief that, if I can mimic, in silicon, a sufficient subset of the fundamental properties of nervous tissue, then I will be able to build computing machines that exhibit behavior analogous to that of nervous systems.

Although my synapse transistors already possess those attributes that I believe are essential for building silicon learning systems, further development will improve the devices. I have already discussed four areas for improvement (see Section 2.3), including reduced tunneling voltages, reduced tunneling-junction leakage, reduced overlap capacitances, and smaller synapse size. In all cases, more modern processing will readily allow these improvements. In addition, as silicon integrated-circuit technology advances, and gate-oxide thickness' scale below 40Å, I will be able to use direct tunneling [4], rather than FN tunneling, to modify the floating-gate charge.

This improvement will allow me to eliminate the tunneling junction entirely, and both to tunnel and to inject electrons directly from the MOS channel to the floating gate.

5.2 Long-Term Learning in Distributed Systems

In a synaptic array, the stored memories are distributed among the node synapses in a fashion prescribed by the array learning rule. In my four-terminal *n*FET array (see Chapter 4), I used an L1 normalization constraint—I held the sum of the row-synapse weight values constant—to prevent unbounded synapse-weight growth. This constraint defined a learning rule, and likewise defines memory storage in the array: Given an input data set and the learning rule, I can describe the memory representation in the array.

I have fabricated, but have not yet tested, other arrays, including a guarded-*p*FET array that uses an L1 constraint, and a four-terminal *n*FET array that uses my colleague Brad Minch's floating-gate MOS translinear-circuit methodology [5] to enforce an L2 constraint (the sum of the squared synapse weights is held constant). These arrays will embody learning rules that are different from that derived in Section 4.2.2, and likewise will distribute their memories across the synaptic elements in a different fashion.

I can implement yet other constraints in synaptic learning arrays. Each array and constraint is likely to give rise to a different learning rule, and to a different distributed-memory representation. The investigation of these rules and representations is a study of neural-network learning in synapse-based silicon systems. Although the rules and representations that I derive may be unlike existing neural-network algorithms, because they derive naturally from the silicon-MOS physics, and from synaptic devices modeled after neurobiology, they offer the potential for building powerful computing systems modeled after neurobiology.

5.3 Neural Computation and Time

Recent neurophysiological evidence [6] suggests that synapses adapt to their presynaptic input continually, signaling relative changes in the input spike rate, rather than communicating the absolute rate. Consequently, neurophysiologists have begun to consider synapses as dynamic structures that adapt to both amplitude and temporal representations in the presynaptic input, and that process information in both the spatial and temporal domains [7]. Likewise, neural-network researchers have begun to incorporate explicit temporal representations into their models [8, 9, 10], to enable these networks to learn temporal associations and sequences, and to predict events.

The study of dendritic function and dendritic signal processing began with the pioneering work of Ramón y Cajal [11], but remained in the background until the late 1950s, when cable

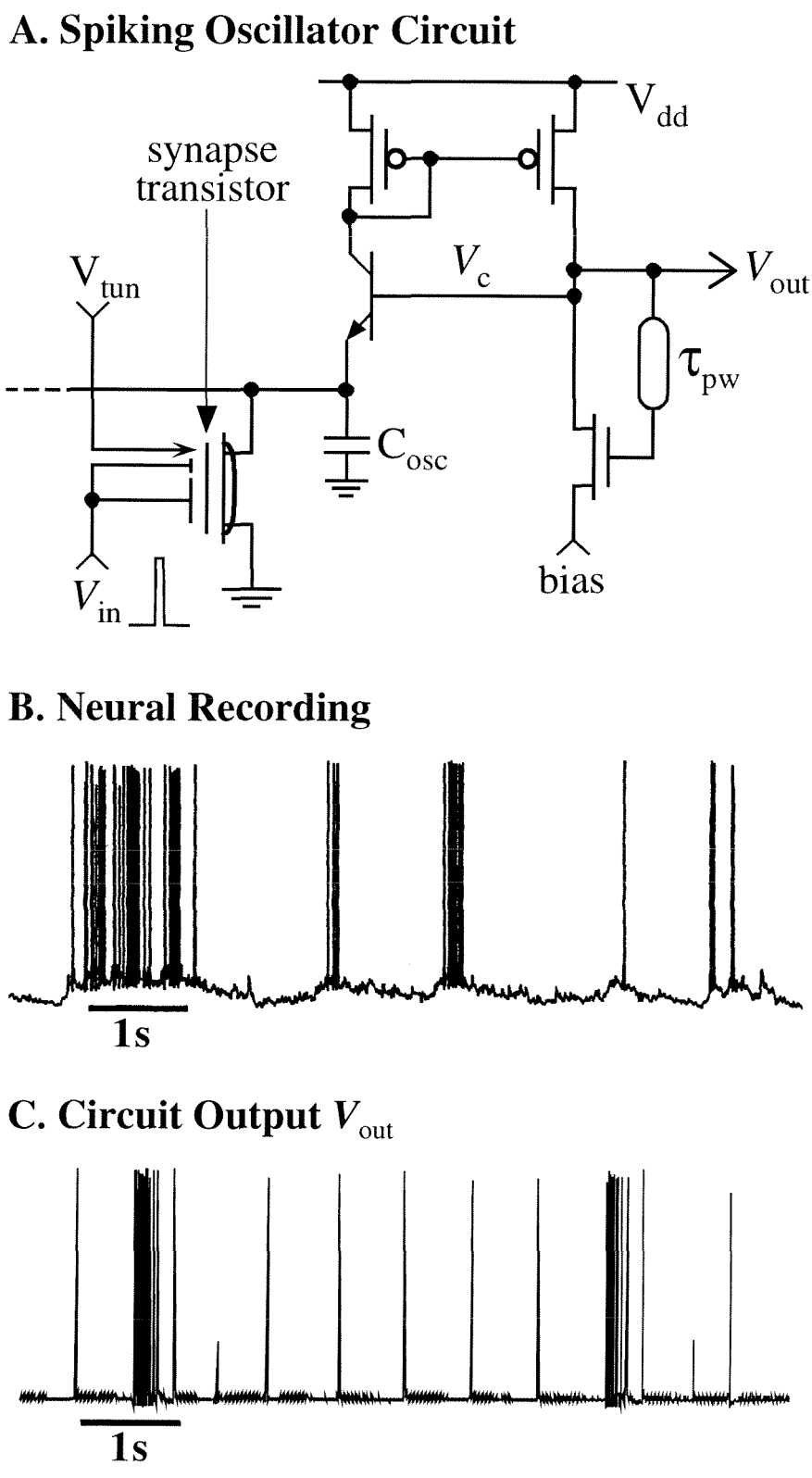


Figure 5.2 Spiking oscillations in neural and silicon synaptic systems. (A) I modified the circuit in Figure 4.2 to induce positive-feedback instability, by removing the integration capacitor from the V_c node and by

adding a destabilizing capacitor C_{osc}. I applied a fixed, high tunneling voltage to the synapse transistor, thereby inducing a small, constant gate current. This gate current caused the synapse transistor's weight to increase gradually; when the weight exceeded a threshold (set by the dynamics of the normalization circuit), the loop became unstable. Instability caused output voltage spikes: V_{out} rose rapidly toward V_{dd}; after a delay τ_{pw}, the rising V_{out} increased the bias transistor's drain current, causing V_{out} to fall rapidly. The width of the output voltage spike was set by the delay τ_{pw}. A rising V_{out} also induced CHEI in the synapse transistor, thereby causing the synapse's weight value to decrease. When the weight value fell below the oscillation threshold, V_{out} remained low until the synapse's weight charged back up. In the absence of external stimuli, the circuit generated voltage spikes at about a 1 Hz rate. I was able to adjust the spike rate over many decades in frequency by changing the tunneling voltage. (B) Intracellular recordings from a cell in the primary somatosensory cortex of the anesthetized cat. The slow oscillations (0.8 Hz to 0.9 Hz) were punctuated by fast, bursting oscillations at a 25 Hz to 40 Hz rate. *Source:* M. Steriade, "Neuromodulatory systems of the thalamus and neocortex," *Seminars in the Neurosciences,* vol. 7, no. 5, pp. 361–370, 1995, © Academic Press Limited, used with permission. (C) I applied a periodic square-wave stimulus to V_{in}, causing large, rapid changes in the synapse's floating-gate voltage. To remove the applied charge, the circuit generated spike bursts coincident with the rising edge of the input voltage pulse V_{in}. I do not imply that my simple silicon circuit actually models the neural system; the waveform in (B) has meaning to the cat's brain, whereas the waveform in (C) is merely the output from a bursting oscillator. However, these data show that floating-gate MOS circuits can exhibit waveforms similar to those seen in neurobiology.

theory and intracellular recording enabled passive dendrite models [12]. Even at that time, however, evidence existed for active properties in dendrites [13]. More recent research (for a review see [14]), indicates that dendrites are active structures whose function can be modulated by activity-dependent mechanisms. More speculative proposals, grounded in neurophysiology, hint at spatiotemporal correlations and spatiotemporal signal processing in dendritic structures (again, see [14]). If correct, these theories and models *implicate dendritic wire as a primary computing element in the brain.* Dendritic wiring may serve not merely to communicate action-potential inputs to the soma, but also to correlate, in space and in time, inputs from vast numbers of neurons, transmitted to the dendrite by time-sensitive and amplitude-sensitive synapses.

The conjecture that the brain is a dynamical system that encodes spatiotemporal information using synaptic weights and dendritic wiring, and that performs massively parallel spatiotemporal correlations using enormous populations of neurons, is both daunting and exciting. Although we cannot yet build integrated circuits that approach the computational density of nervous tissue, with our present silicon technology we can build integrated circuits to investigate the spatiotemporal signal processing that may underlie neuronal computation.

5.4 Closing Remarks

John von Neumann, in 1945 [15], introduced the computational paradigm that forms the basis for nearly all machine computation to date. Contemporary digital computers are amazingly powerful machines, and far outperform the brain on tasks that can be described mathematically. However, when confronted with an ill-posed problem, such as identifying a tree in a visual field, these same digital computers fail miserably. By contrast, the brain excels at quickly finding good solutions to ill-posed problems. Neurobiology demonstrates—by means of working examples—an alternative computational paradigm to the von Neumann machine. The notion that alternative computing paradigms not only exist, but also can form computing machines optimized for solving different kinds of problems, is my real motivation for studying neural computation.

Neural computation presents us with a wealth of unknowns. The mechanisms of neuronal signaling, communication, development, and learning are unknown. The general anatomical loci of thoughts, concepts, memories, and emotions is unknown. From the point of view of synaptic plasticity, even the microstructure of representations remains unknown. Because of the brain's large size (in information-theoretic terms), and the current paucity of theories describing complex dynamical systems, we cannot hope to understand neural computation by observation or by experiment alone. Instead, we must construct experiments that interact with theory, to produce testable hypothesis' and predictions [3]. By developing neurally inspired learning systems, in our advanced silicon technology, I hope to shed light on the computational paradigm that neurobiology uses to solve real-world problems.

References

1 C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A high-resolution nonvolatile analog memory cell," in *Proc. IEEE Intl. Symp. on Circuits and Systems*, Seattle, WA, vol. 3, pp. 2233–2236, 1995.

2 P. Hasler, B. A. Minch, C. Diorio, and C. Mead, "An autozeroing amplifier using *p*FET hot-electron injection," in *Proc. IEEE Intl. Symp.* on *Circuits and Systems*, Atlanta, GA, vol. 3, pp. 325–328, 1996.

3 K. Jeffery and I. Reid, "Modifiable neuronal connections: An overview for psychiatrists," *Am. J. of Psychiatry,* vol. 154, no. 2, pp. 156–164, 1997.

4 E. Takeda, C. Yang, and A. Miura-Hamada, *Hot-Carrier Effects in MOS Devices,* San Diego, CA: Academic Press, 1995.

5 B. Minch, *Analysis, Synthesis, and Implementation of Networks of Multiple-Input Translinear Elements,* Ph.D. Thesis, Department of Computation and Neural Systems, California Institute of Technology, Pasadena, CA, 1997.

6 H. Markram and M. Tsodyks, "Redistribution of synaptic efficacy between neocortical pyramidal neurons," *Nature,* vol. 382, no. 6594, pp. 807–810, 1996.

7 C. Koch, "Computation and the single neuron," *Nature,* vol. 385, no. 6613, pp. 207–210, 1997.

8 P. Montague, P. Dayan, C. Person, and T. Sejnowski, "Bee foraging in uncertain environments using predictive Hebbian learning," *Nature,* vol. 377, no. 6551, pp. 725–728, 1995.

9 J. J. Hopfield, "Transforming neural computations and representing time," *P. Natl. Acad. Sci.,* vol. 93, no. 26, pp. 15440–15444, 1996.

10 J. J. Hopfield, "Pattern recognition computation using action potential timing for stimulus representation," *Nature,* vol. 376, no. 6535, pp. 33–36, 1995.

11 S. Ramón y Cajal, *Histology of the Nervous System of Man and Vertebrates,* vol. 2, translated from the French version of the original Spanish by N. Swanson and L. W. Swanson, New York: Oxford University Press, 1995.

12 W. Rall, *The Theoretical Foundations of Dendritic Function,* I. Segev, J. Rinzel, and G. M. Shepherd, eds., Cambridge: MIT Press, 1995.

13 R. Lorente de Nó and G. A. Coundouris, "Decremental conduction in peripheral nerve integration of stimuli in the neuron," *P. Natl. Acad. Sci.,* vol. 45, pp. 592–617, 1959.

14 R. Yuste and D. W. Tank, "Dendritic integration in mammalian neurons, a century after Cajal," *Neuron,* vol. 16, no. 4, pp. 710–716, 1996.

15 J. von Neumann (1945/1982), "First draft of a report on the EDVAC," in *The Origins of Digital Computers: Selected Papers,* B. Randall ed., 3rd edition, Berlin: Springer-Verlag, 1982.