Chapter 3 Single-Transistor Synapses

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

3.1 Overview of Single-Transistor Learning Synapses

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

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

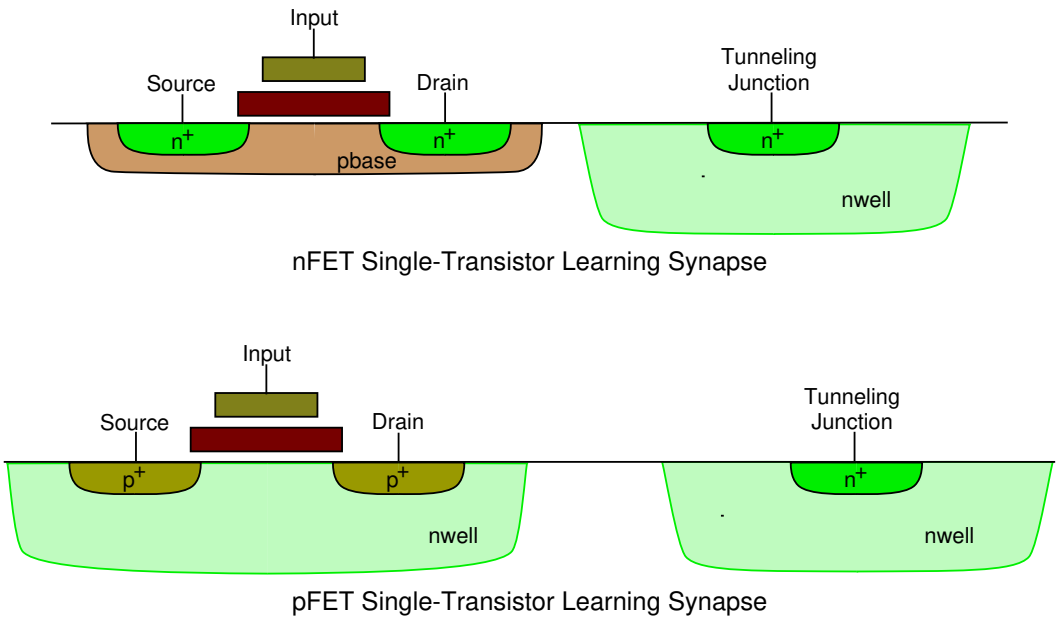


Figure 3.1: Cross section of the nFET and pFET single-transistor synapses in an $n$well MOSIS process. The tunneling junctions used by the single-transistor synapses is a region of gate oxide between the polysilicon floating-gate and $n$well. For the nFET synapse, the pbase implant results in a larger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

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

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

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

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

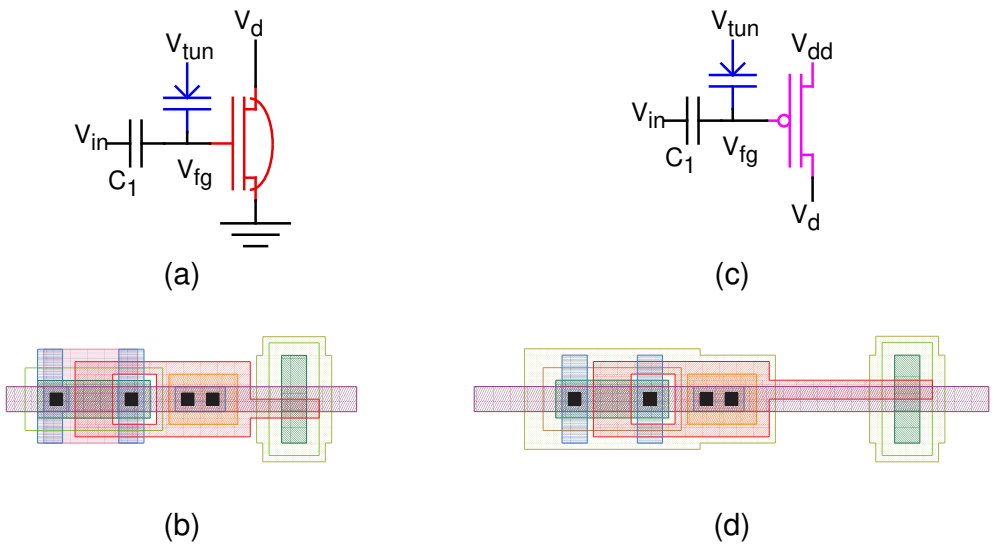


Figure 3.2: Circuit symbols of the nFET and pFET single-transistor synapses. (a) Circuit diagram of the nFET single-transistor synapse with its source connected to ground. (b) Layout of the nFET single-transistor synapse. (c) Circuit diagram of the pFET single-transistor synapse with its source connected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

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

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

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

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

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

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

3.2 Nonadaptive Behavior

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

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

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

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

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

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

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

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

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

Neglecting Early voltage effects, the source current becomes

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

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

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

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

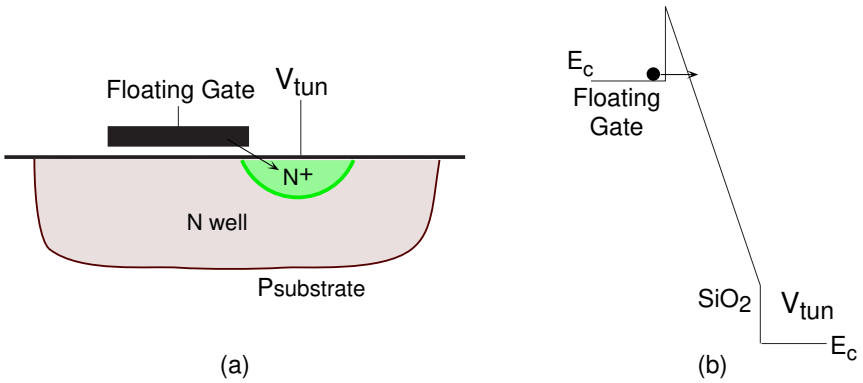


Figure 3.3: Tunneling in an $n$well process. (a) The tunneling junction is the capacitor between the floating gate and the $n$well; we use high-quality gate oxide to reduce the effects of electron trapping. Over a wide range of oxide voltage, most of the tunneling occurs between the floating gate and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

behavior at the fast timescales as

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

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

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

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

3.3 Electron Tunneling

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

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

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

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

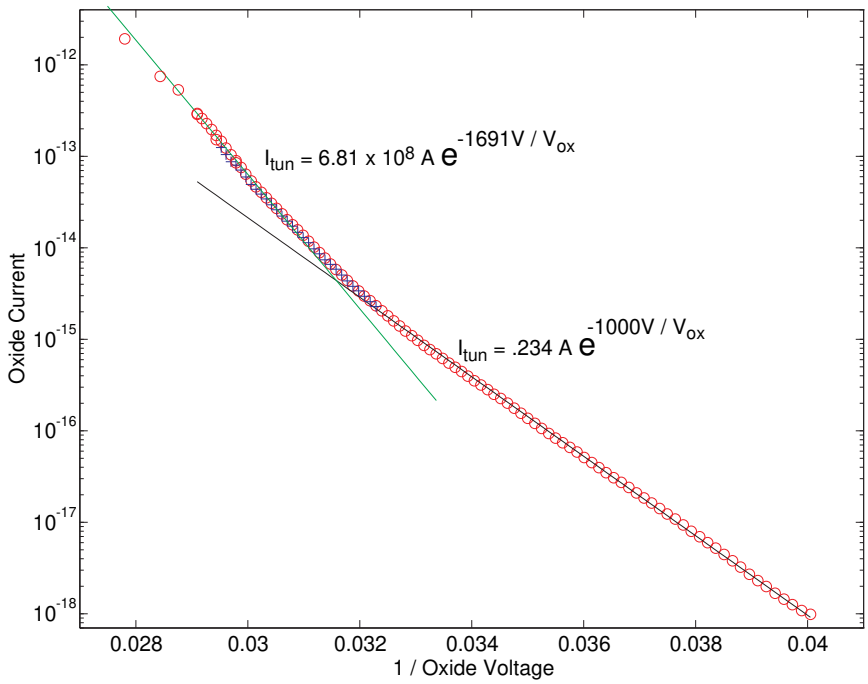


Figure 3.4: Electron tunneling current versus 1/oxide voltage. The two straight line fits are to the classic Fowler-Nordheim expression in (3.8). The two different straight-line regions might be due to tunneling through intermediate traps, or due to initially tunneling through the junction edge for low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

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

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

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

where

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

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

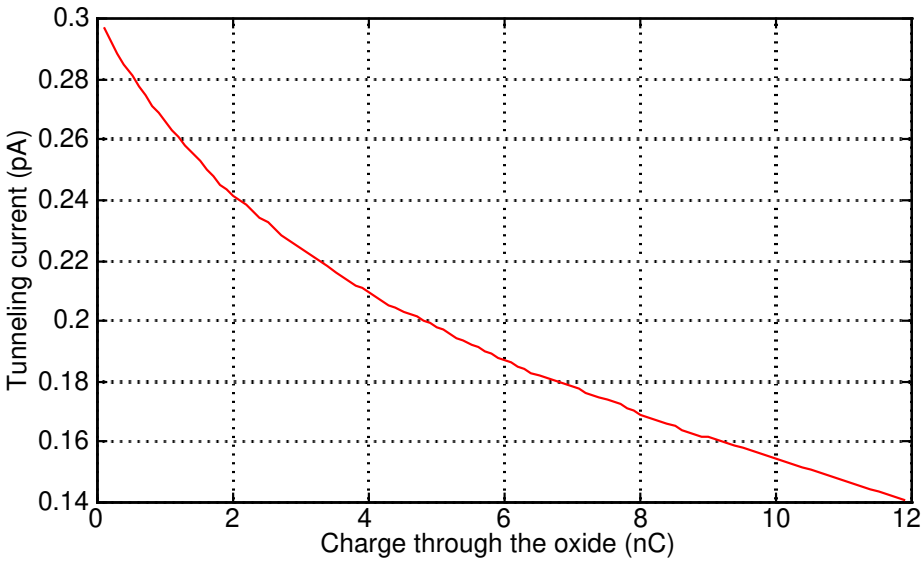


Figure 3.5: Electron tunneling current at a fixed oxide voltage ($V_{ox} = 33$V) versus charge through our tunneling junction. This experiment induced 60V of charge on a 200pF capacitor; to first order, tunneling trap creation is a function of the total charge through the oxide. This measurement characterizes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

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

$$\begin{aligned} n\text{FET}: I_{tun} &= I_{tun0}\left(\frac{I_s}{I_{so}}\right)^{-\frac{U_T}{\kappa_n V_x}}, \\ p\text{FET}: I_{tun} &= I_{tun0}\left(\frac{I_s}{I_{so}}\right)^{\frac{U_T}{\kappa_p V_x}}. \end{aligned} \tag{3.11}$$

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

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

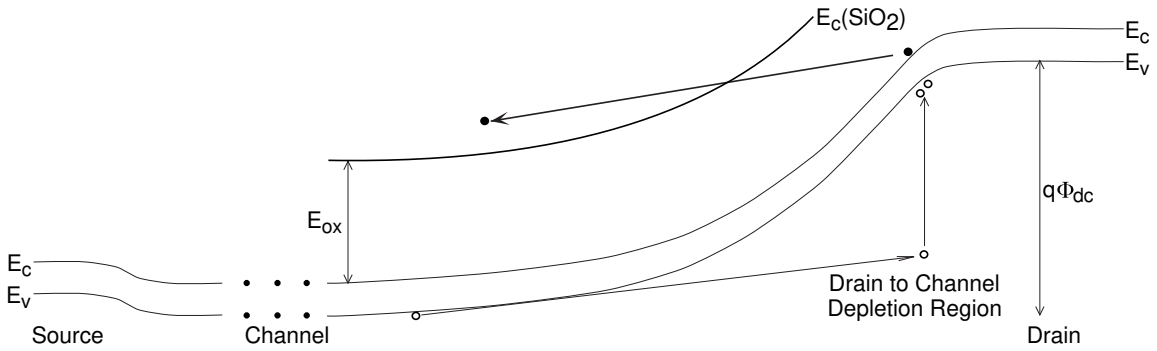


Figure 3.6: Band diagram of a subthreshold pFET transistor under conditions favorable for hot-electron injection. The source of electrons to be injected to the floating-gate is created by hole impact ionization. E_{ox} is the Si–SiO_2 barrier, which is 3.04eV for no field across the oxide.

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

3.3.1 Hot-Electron Injection

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

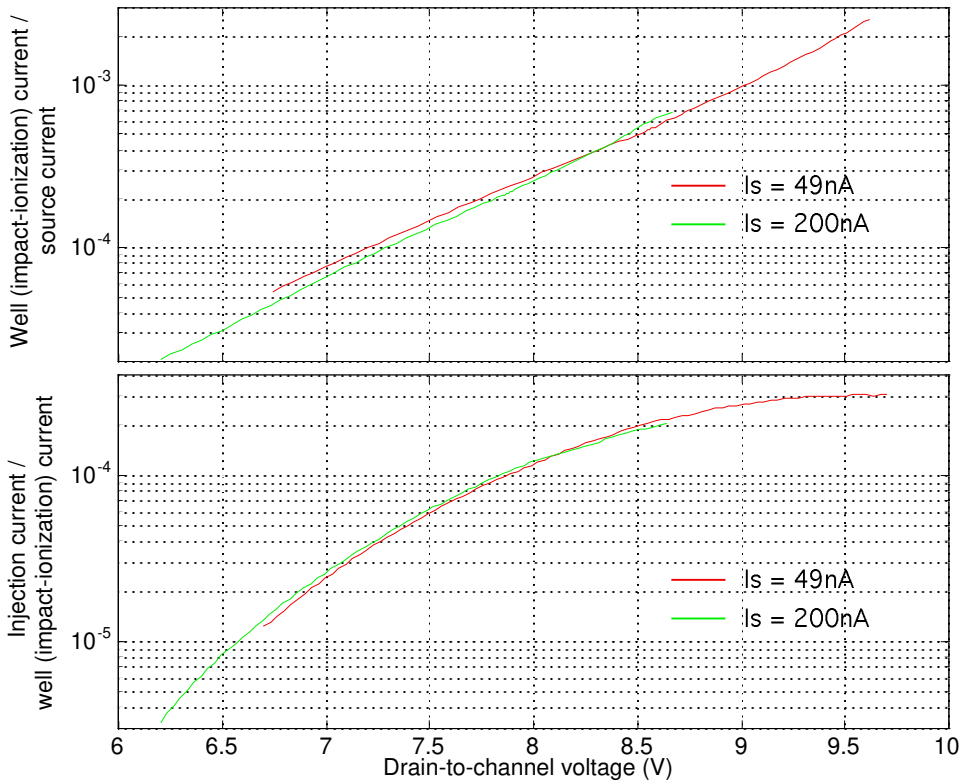


Figure 3.7: Measured data of the current dependences in hot-electron injecting pFETs versus the drain-to-channel voltage for two source currents. The top plot shows the ratio of impact-ionization current and source current versus drain-to-channel potential. The lower plot shows the ratio of hot-electron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

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

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

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

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

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

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

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

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

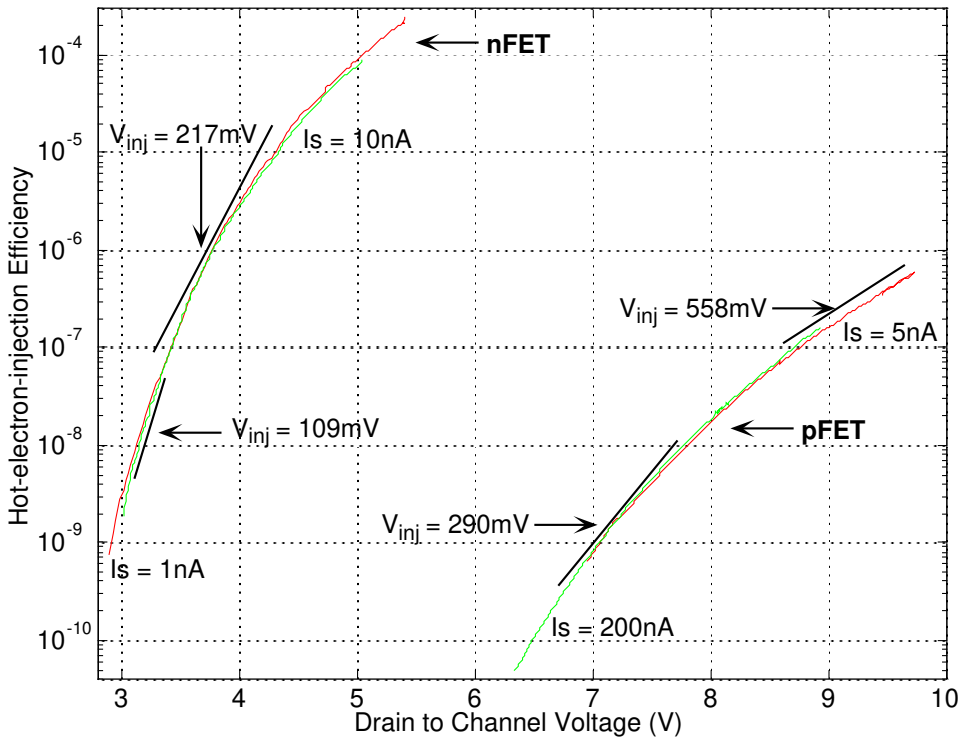


Figure 3.8: nFET and pFET hot-electron injection efficiency ($\frac{I_{inj}}{I_s}$) versus Φ_{dc} for two values of source current. Injection efficiency is the ratio of injection current to source current. The two different source current values are nearly equal, which is consistent with injection efficiency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

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

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

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

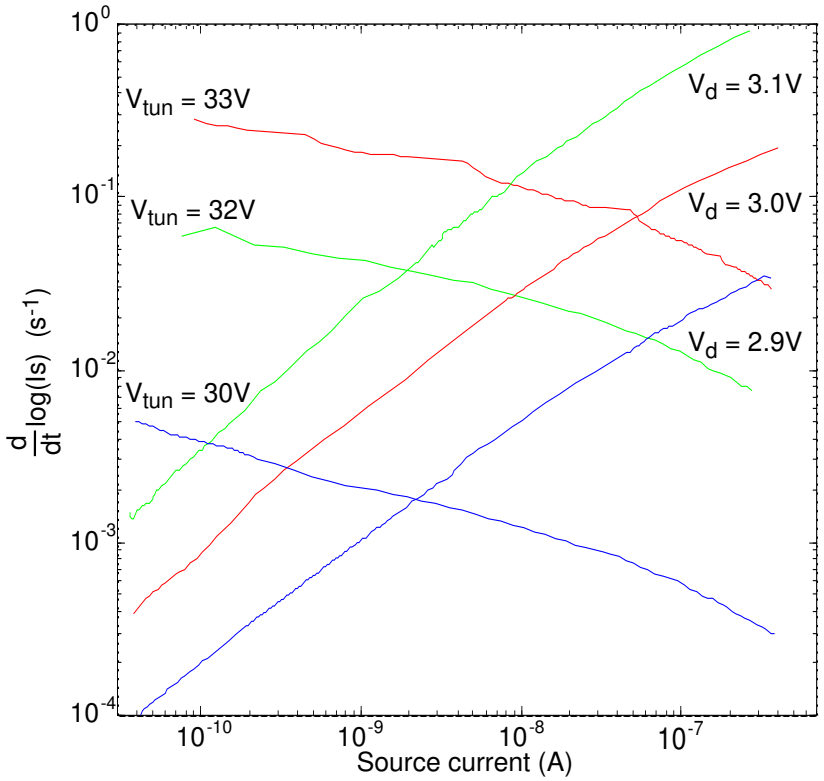


Figure 3.9: Plot of $\frac{d}{dt}\log(I_s)$ as a function of I_s for three different tunneling and three different drain voltages. Starting at an appropriately low (or high) source current (I_s), I measured this data by stepping to the desired drain or tunneling voltage, and letting the tunneling (or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

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

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

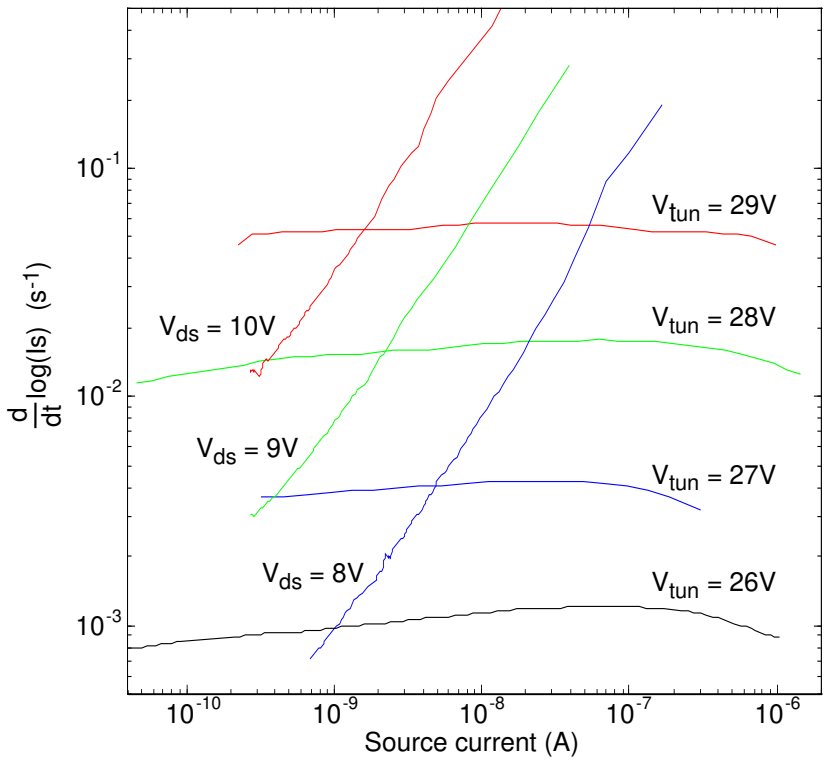


Figure 3.10: Plot of $\frac{d}{dt}\log(I_s)$ as a function of I_s for four different tunneling and three different drain voltages. Plotting the data in this way shows the power-law dependence of channel current on the floating-gate current. I measured this injection (tunneling) data as in Fig. 3.9 by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

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

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

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

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

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

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

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

3.4 An Array of Single Transistor Synapses

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

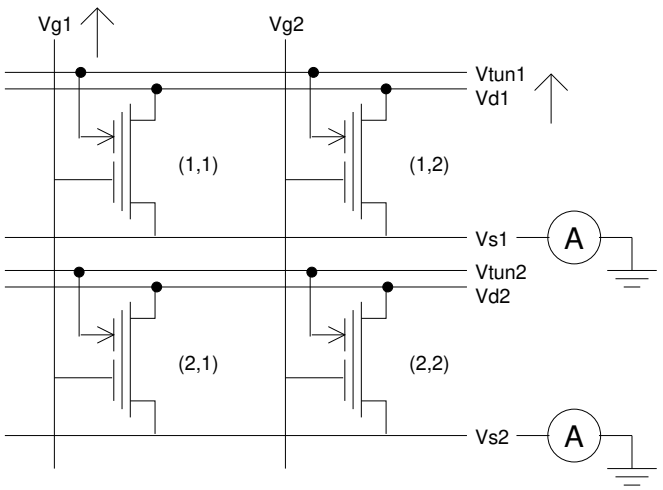


Figure 3.11: Circuit diagram of the single-transistor synapse array. Each transistor has a floating gate capacitively coupled to an input column line. A 2 x 2 section of the array allows us to characterize how modifying a single floating gate (such as synapse (1,1)) affects the neighboring floating gate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

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

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

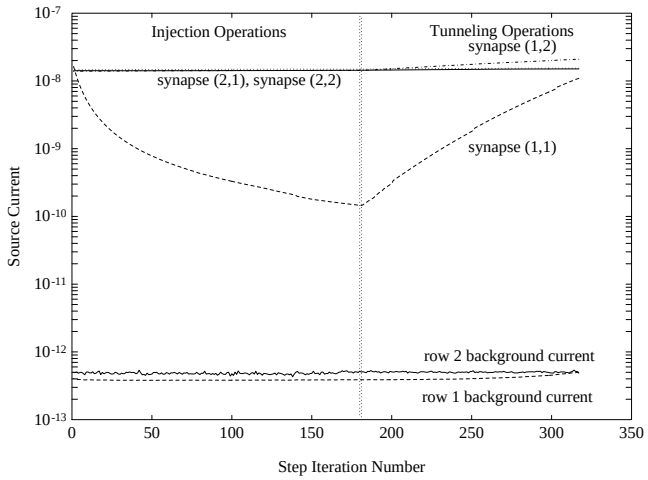
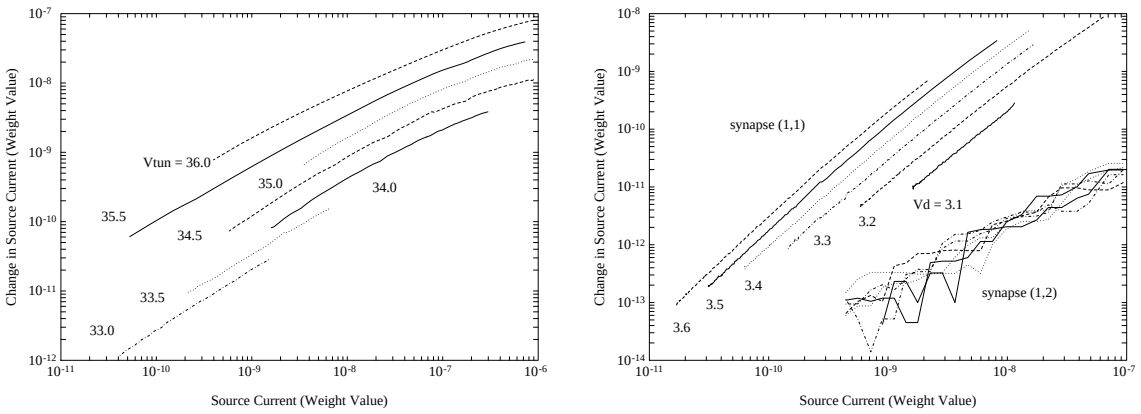


Figure 3.12: Output currents from a 2 x 2 section of the synapse array, showing 180 injection operations followed by 160 tunneling operations. Because our measurements from the 2 x 2 section come from a larger array, we also display the 'background' current from all other synapses on the row. This background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

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

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



(a) (b)

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

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

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

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

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

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

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

and for the pFET synapse as

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

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

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