Chapter 6 Conclusions and Future Directions

This dissertation described the foundations of a floating-gate technology for adaptation and learning. I developed an analytical model of hot-electron injection and impact ionization from first principles, and described supportive experimental measurements from subthreshold nFETs and pFETs. I described the invention, characterization, and modeling of the single-transistor learning synapses. Then, I characterized and modeled the continuous-time floating-gate dynamics of simple circuits of single-transistor learning synapses. Finally, I described the invention, characterization, and modeling of the autozeroing floating-gate amplifier. These four building blocks form the foundation for several large-scale floating-gate adaptive systems.

6.1 Summary of Thesis Results

Chapter 2 presented a model hot-electron transport in the drain-to-channel depletion region using the spatially varying Boltzmann transport equation, and analytically found a self-consistent distribution function in a two-step process. In the first step, I solved for the average hot-electron trajectory in energy and direction as a function of position through the depletion region. In the second step, I solved for the electron distribution function around this average electron trajectory. In this coordinate system, phonon collisions spatially diffuse the distribution function, and impact-ionization collisions remove electrons from the high-energy distribution function. From the electron distribution function, I calculated the probabilities of impact ionization and hot-electron injection as functions of channel current, drain voltage, and floating-gate voltage. I compared the results of my analytical model to measurements in long-channel devices. The model simultaneously fits both the hot-electron-injection and impact-ionization data. These analytical results yield an energy-dependent impact-ionization collision rate that is consistent with numerically calculated collision rates reported in the literature.

To make the transition from devices to circuits, Chapter 3 presented the single-transistor learning synapses. The single-transistor synapse, invented and developed by me and my collaborators, Diorio and Minch, is an hot-electron injecting FET with a well-tunneling junction. The single-transistor synapses simultaneously perform long-term weight storage, compute the product of the input and the weight value, and update the weight value according to a Hebbian or a backpropagation learning rule. Memory is accomplished via charge storage on polysilicon floating gates, providing long-term retention without refresh. The small size and low power operation of single-transistor synapses allows the development of dense synaptic arrays. The synapses efficiently use the physics of silicon to perform weight updates; electron tunneling increases the weight value and hot-electron injection decreases the weight value. I presented a model of electron tunneling, which I use to remove electrons from the floating gate, and I presented hot-electron-injection measurements in both nFET and pFETs. The charge on the floating gate is decreased by hot-electron injection with high selectivity for a particular synapse. The charge on the floating gate is increased by electron tunneling, which results in high selectivity between rows, but in much lower selectivity between columns along a row.

Chapter 4 presented the possible negative- and positive-feedback configurations of continuous-time floating- gate MOS circuits. The learning synapses provide different floating-gate dynamics depending on their configuration. I built a new type of pFET synapse by degenerating the source; the oxide currents of the synapse provide stabilizing feedback to the floating gate and to the drain. I discussed the range of possible stabilizing and destabilizing types of feedback in circuits with one floating-gate synapse, including data from nFET, pFET, and source-degenerated pFET synapses. Multiple floating-gate circuits show competitive and cooperative behavior between synapses, and that is consistent with Hebbian and anti-Hebbian learning. Modeling the dynamics of a floating-gate circuit builds intuition to develop more complex circuits, and shows the strengths of simple floating-gate current models.

The autozeroing floating-gate amplifier (AFGA) is a simple example of this class of adaptive floating-gate MOS circuits; these circuits use tunneling and hot-electron injection to adapt the charge on floating gates to return the circuit to a baseline condition on a slow timescale. When the appropriate feedback is applied to the floating gate, this adaptation is an inherent part of the circuit's operation—no additional control circuitry is required. The modulation of the pFET hot-electron injection by the output voltage provides the correct feedback to return the output voltage to the proper operating regime.

Chapter 5 presented a bandpass floating-gate amplifier that uses tunneling and pFET hot-electron injection to set its DC operating point adaptively. Because the hot-electron injection is an inherent part of the pFETs behavior, I obtain this adaptation with no additional circuitry. Because the gate currents are small, the circuit exhibits a high-pass characteristic with a cutoff frequency less than 1 Hz. The high-frequency cutoff is controlled electronically, as is done in continuous-time filters. I have derived analytical models that completely characterize the amplifier and that are in good agreement with experimental data for a wide range of operating conditions and input waveforms. The AFGA always is a first-order system, even in the presence of parasitic capacitances; therefore, the AFGA is unconditionally stable, with 90 degrees of phase margin for noninductive loads. Since the AFGA's noise performance is similar in thermal and $1/f$ characteristics to that of a standard MOS amplifier, the tunneling and injection processes do not add appreciable noise to the amplifier. In addition, with a desired adaptation rate, I can reduce significantly the low-frequency noise generated in the AFGA. I can increase the linear range by increasing C_w, and I can increase the dynamic range by increasing C_w or C_L. This autozeroing floating-gate amplifier demonstrates how to use continuous-time, floating-gate adaptation in amplifier design.

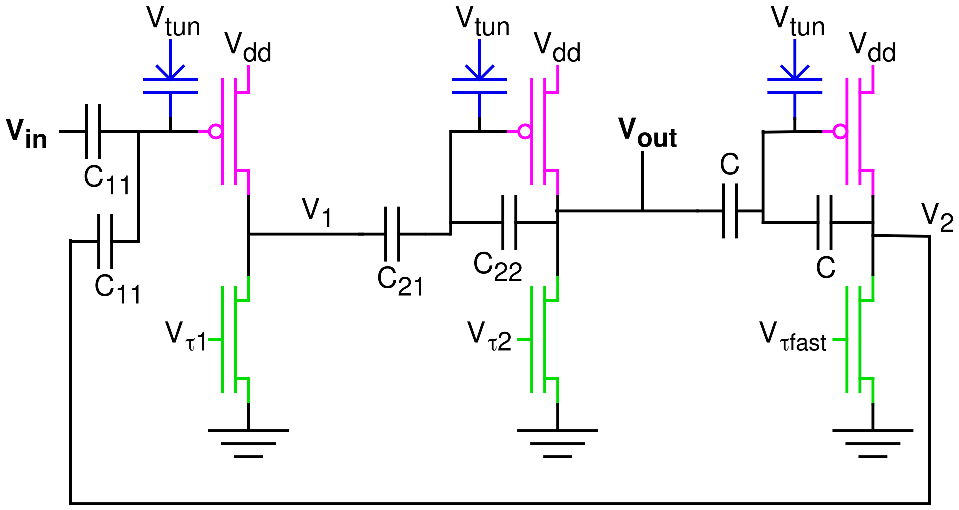


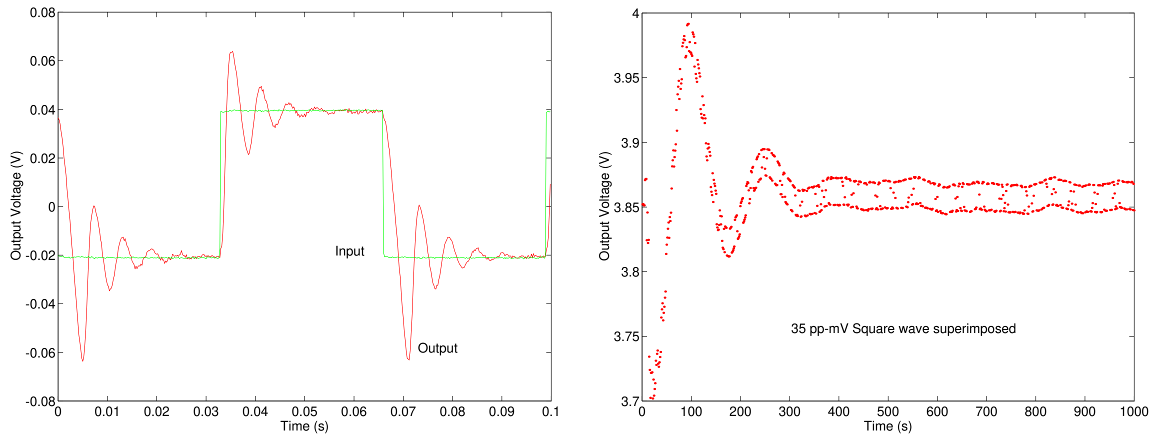
Figure 6.1: Circuit diagram of the autozeroing second-order section. This circuit, which is built with three autozeroing amplifiers, shows second-order behavior that is electronically controlled.

6.2 Future Directions

This thesis considered the spectrum of a silicon floating-gate technology for adaptation and learning spanning from hot-electron transport to simple floating-gate circuits and dynamics. The remaining question is whether this technology will scale to large floating-gate circuits and systems with adaptation and learning. The rest of this section briefly presents two floating-gate circuits, the autozeroing second-order section and the adaptive winner-take-all; these circuits show that this technology can scale to larger circuits and systems.

6.2.1 Autozeroing Second-Order Section

Figure 6.1 shows my autozeroing second-order section; it is the primary element to build higher-order filters, both in continuous-time and switch-capacitor circuits. This circuit is based on the Diff2 second-order section [82], which in turn is based on a canonical form for a second-order section as a high-gain amplifier with feedback. To build a simplified model of the second-order behavior, we assume that $V_2 = -V_{out}$; that is, $V_{\tau fast}$ is set at a bias current much larger than $V_{\tau 1}$ or $V_{\tau 2}$. We ignore (for



(a) (b)

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

the moment) the floating-gate currents, ignore the capacitive feedthrough effects, and assume the $V_{\tau 1}$ amplifier's gain is very large. From the AFGA analysis in Chapter 5, the small-signal model of the two amplifiers is

$$\frac{C_{T1}C_{o1} - C_{21}^2}{C_{21}} \frac{U_T}{\kappa I_{\tau 1}} \frac{dV_1}{dt} = (V_{in} + V_2)\,, \tag{6.1}$$

$$\frac{C_{T2}C_{o2} - C_{22}^2}{C_{22}} \frac{U_T}{\kappa I_{\tau 2}} \frac{dV_{out}}{dt} = (V_1 + V_{out})\,, \tag{6.2}$$

where C_T is the total capacitance connected to the floating gate ($C_T = C_1 + C_2 + C_w$), C_o is the total capacitance connected to the output node ($C_o = C_2 + C_L$), and C_{21}, C_{22} are the feedback capacitors from floating gate to output. By taking the Laplace transform of these equations, we solve for the transfer function from V_{in} to V_{out} as

$$\frac{V_{out}(s)}{V_{in}(s)} = \frac{1}{1 + s\frac{C_{T1}C_{o1} - C_{21}^2}{C_{21}}\frac{U_T}{\kappa I_{\tau 1}} + s^2 \frac{C_{T1}C_{o1} - C_{21}^2}{C_{21}}\frac{CU_T}{\kappa I_{\tau 1}}\frac{C_{T2}C_{o2} - C_{22}^2}{C_{22}}\frac{CU_T}{\kappa I_{\tau 2}}} = \frac{1}{1 + s\tau_1 + s^2\tau_1\tau_2}, \tag{6.3}$$

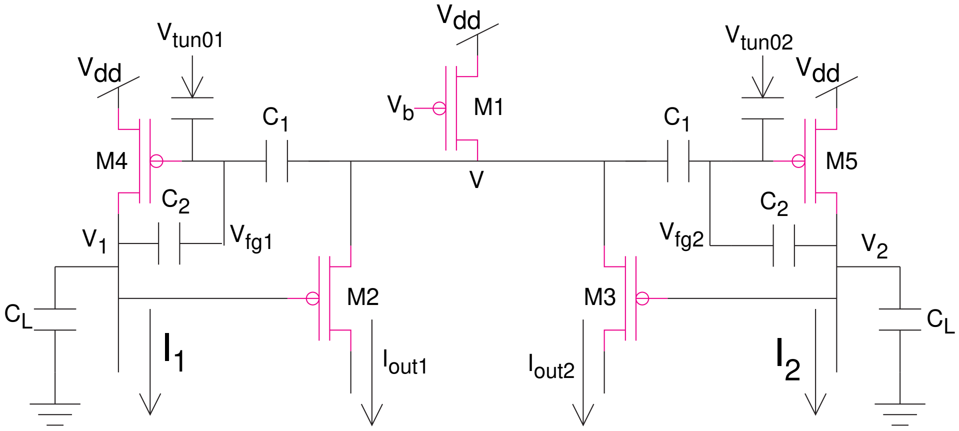


Figure 6.3: The circuit diagram of a two-input winner-take-all circuit.

where I define

$$\tau_1 = \frac{C_{T1}C_{o1} - C_{21}^2}{C_{21}}\frac{U_T}{\kappa I_{\tau 1}}, \text{and } \tau_2 = \frac{C_{T2}C_{o2} - C_{22}^2}{C_{22}}\frac{U_T}{\kappa I_{\tau 2}}. \tag{6.4}$$

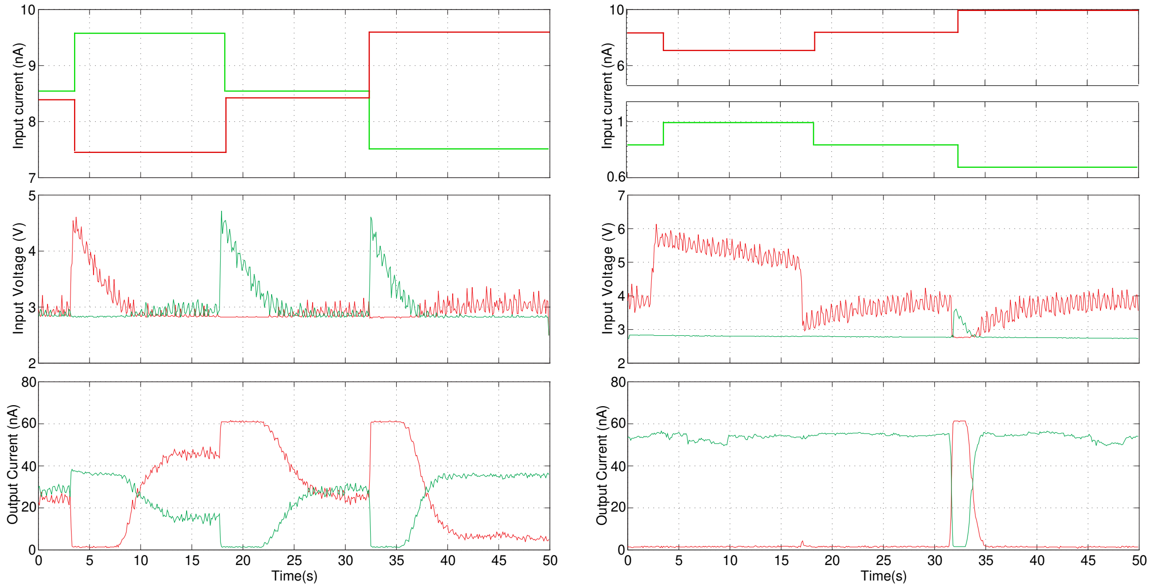
From the canonical form for the second-order section, we find that [82]

$$\tau = \sqrt{\tau_1 \tau_2}, \text{and } Q = \sqrt{\frac{\tau_1}{\tau_2}}. \tag{6.5}$$

Figure 6.2(a) shows the step response of this autozeroing second-order section. The data show the characteristic ringing behavior of a second-order system; changing $V_{\tau 1}$ and $V_{\tau 2}$ changes τ and Q, as predicted from (6.5). Figure 6.2(b) shows the step response at a much slower timescale. The oscillatory behavior is due to the corner frequencies, which are set by the floating-gate currents. This part of the dynamics is usually third order, and therefore the modeling becomes more difficult.

6.2.2 Adaptive Winner-Take-All

Figure 6.3 shows the two-input, adaptive winner–take–all (WTA) circuit, invented by W. Fritz Kruger and myself [93], based on Lazzaro's classic WTA [92]. We have added a time dimension (adaptation) to make the input derivative an important factor in



(a) (b)

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

winner selection. The difference between the classic and adaptive WTA is that $\boldsymbol{M_4}$ and $\boldsymbol{M_5}$ are pFET single-transistor synapses; these floating-gate transistors null their inputs slowly over time. This enhancement results in the ability of each transistor to adapt to its input bias current. The adaptation is a result of the electron tunneling and hot-electron injection modifying the charge on the floating gate. The circuit is devised such that these are negative-feedback mechanisms; consequently, the output voltage always returns to the same steady-state voltage, determined by its bias current regardless of the DC input level.

Figure 6.4 shows characteristic traces from the two-input circuit. The winning node corresponds to the lowest voltage, which is reflected in that node's corresponding high output current. Looking at Fig. 6.4(a) we see that, as an input step is applied, the output current jumps and then begins to adapt to a steady-state value. When the inputs are nearly equal, the steady-state outputs are nearly equal; when the inputs are different, the steady state output is greater for the cell with the lesser input. In general, the input current change that is the largest after reaching the previous equilibrium becomes the new equilibrium. This additional decrease in V_1 would lead to an amplified increase in the other voltage since the losing stage roughly looks like an autozeroing amplifier with the common node as the input terminal. The extent to which the inputs do not equal this largest input is manifested as a proportionally larger input voltage. The other voltage would return to equilibrium by slowly, linearly decreasing in voltage due to the tunneling current. This process will continue until V_1 equals V_2. Note that in general that the inputs with lower bias currents have a slight starting advantage over the inputs with higher bias currents.

Figure 6.4(b) illustrates the advantage of the adaptive WTA over the classic WTA. In the classic WTA, the output voltage and current would not change throughout the experiment, but the adaptive WTA responds to changes in the input. The second input step does not evoke a response because there was not enough time to adapt to steady state after the previous step; but the next step immediately causes it to win.