Adaptive Circuits Using pFET Floating-Gate Devices

Paul Hasler
Georgia Institute of Technology
Atlanta, GA 30332-0250
phasler@ee.gatech.edu

Bradley A. Minch
Cornell University
Ithaca, NY 14853-5401
minch@anise.ee.cornell.edu

Chris Diorio
University of Washington
Seattle, WA 98195-2350
diorio@cs.uwashington.edu

Abstract

In this paper, we describe our floating-gate pFET device, with its many circuit applications and supporting experimental measurements. We developed these devices in standard double-poly CMOS technologies by utilizing many effects inherent in these processes. We add floating-gate charge by electron tunneling, and we remove floating-gate charge by hot-electron injection. With this floating-gate technology, we can not only build analog EEPROMs, we can also implement adaptation and learning when we consider floating-gate devices to be circuit elements with important time-domain dynamics. We start by discussing non-adaptive properties of floating-gate devices and we present two representative non-adaptive applications. First, we discuss using the floating-gate pFETs as non-volatile voltage sources or potentiometers (e-pots). Second, we will discuss using floating-gate pFETs to build translinear circuits that compute the product of powers of the input currents. We then discuss the physics, behavior, and applications of adaptation using floating-gate pFETs. The physics of adaptation start with floating-gate pFETs with continuous tunneling and injection currents. A single floating-gate MOS device operating with continuous-time tunneling and injection currents can exhibit either stabilizing or destabilizing behaviors. One particular application is an autozeroing floating-gate amplifier (AFGA) that uses tunneling and pFET hot-electron injection to adaptively set its DC operating point. Continuous-time circuits comprising multiple floating-gate MOS devices show various competitive and cooperative behaviors between devices. These floating-gate circuits can be used to build silicon systems that adapt and learn.

In 1967, Kahng and Sze reported the first floating-gate structure as a mechanism for nonvolatile information storage [1]. Since then, floating-gate transistors have been used widely to store digital information for long periods in structures such as EPROMs, EEPROMs, and flash memories. These digital nonvolatile memory technologies have only been fabricated on specialized and expensive processes. FGMOS transistors have also been used as long-term non-volatile information storage devices for analog applications [2].

The focus of this paper is show that floating-gate devices are not just for memories anymore, but are also computational circuit elements with analog memory and important time-domain dynamics. We will paint a coherent picture of the capabilities of pFET floating-gate devices starting from the work that we [3-15] and other colleagues [16-22] have developed over the last 10 years, and including recent research results. We emphasize pFET floating-gate devices in this paper because any CMOS process will always have pFET injection for subthreshold and above-threshold biases. A second and related focus is to show the availability and functionality of these floating-gate devices on standard CMOS processes. To build floating-gate devices in a standard CMOS process, we use several inherent physical mechanisms (hot-electron injection, electron tunneling, DIBL) that are avoided by classical IC designers, mechanisms that will become more pronounced as the line-widths of CMOS processes continue to scale down. All figures in this paper are our experimental measurements from floating-gate devices fabricated in 2.0μm and 1.2μm MOSIS processes.

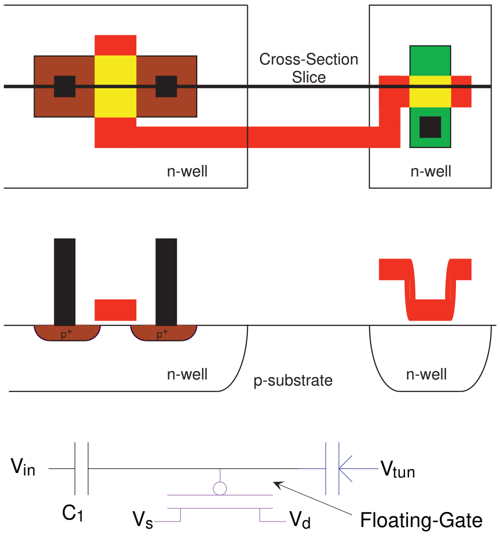


Figure 1: Layout, cross section, and circuit diagram of the floating-gate pFET in a standard double-poly nwell MOSIS process. The cross section corresponds to the horizonatal line slicing through the layout view. The pFET transistor is the standard pFET transistor in the nwell process. The gate input capacitively couples to the floating-gate by either a poly-poly capacitor, a diffused linear capacitor, or a MOS capacitor, as seen in the circuit diagram (not explicitly shown in the other two figures). We add floating-gate charge by electron tunneling, and we remove floating-gate charge by hot-electron injection. The tunneling junctions used by the single-transistor synapses is a region of gate oxide between the polysilicon floating-gate and nwell (a MOS capacitor). Between V_{tun} and the floating-gate is our symbol for a tunneling junction, a capacitor with an added arrow designating the charge flow.

1 Basic Floating-Gate Devices

Figure 1 shows the layout, cross-section, and circuit symbol for our floating-gate pFET device. A floating gate is a polysilicon gate surrounded by SiO_2. Charge on the floating gate is stored permanently, providing a long-term memory, because it is completely surrounded by a high-quality insulator. From the layout, we see that the floating-gate is a polysilicon layer that has no contacts to other layers; this floating-gate can be the gate of a MOSFET and can be capacitively connected to other layers. In circuit terms, a floating-gate occurs when we have no DC path to a fixed potential, precisely the effect avoided by many circuit designers and circuit simulators. No DC path implies only capacitive connections to the floating node, as seen in Fig. 1.

We will focus our discussions on pFETs operating with subthreshold channel currents. Many of the behaviors extend qualitatively to above-threshold operation [5], the quantitative behaviors do not. Because we are considering signals that are capacitively coupled to the floating gate, we model voltage and current swings around the circuit's steady-state values. We describe the subthreshold pFET channel current in saturation, I_s, around a bias current, I_{so}, for a change in the FET's floating-gate voltage, ΔV_{fg}, and source voltage, ΔV_s, as [23]

$$I_s = I_{so} \exp\left(\frac{\Delta V_s - \kappa \Delta V_{fg}}{U_T}\right), \tag{1}$$

where κ is the fractional change in the pFET surface potential due to a change in floating-gate voltage, and U_T is the thermal voltage, $\frac{kT}{q}$. Figure 2a shows experimental measurements from a pFET, where gate voltage is measured down from its substrate (V_{dd}) voltage. The data closely agree with (1) in the subthreshold region, and is also the baseline case for the floating-gate devices.

Figure 2a shows experimental measurements from a basic pFET devices, and from two floating-gate pFETs with different capacitive couplings into the floating-gate. Capacitive coupling from an input into the floating-gate sets up a capacitive-voltage divider from the input to the floating-gate. The fractional charge that reaches the floating-gate is the ratio between the coupling capacitor and the total capacitance, C_T, connected to the floating gate. In Fig. 2a, the exponential slopes are smaller than the basic pFET device, as expected from the capacitive voltage divider. Also, the exponential slope when coupling with two identical capacitors was twice that when coupling with only a single capacitor. One can explicitly design various different exponential and logarithmic functions, where the accuracy of the ratio of these exponential slopes is only dependent on the quality of capacitor matching in a given process [11]. We can modify (1) to model the current

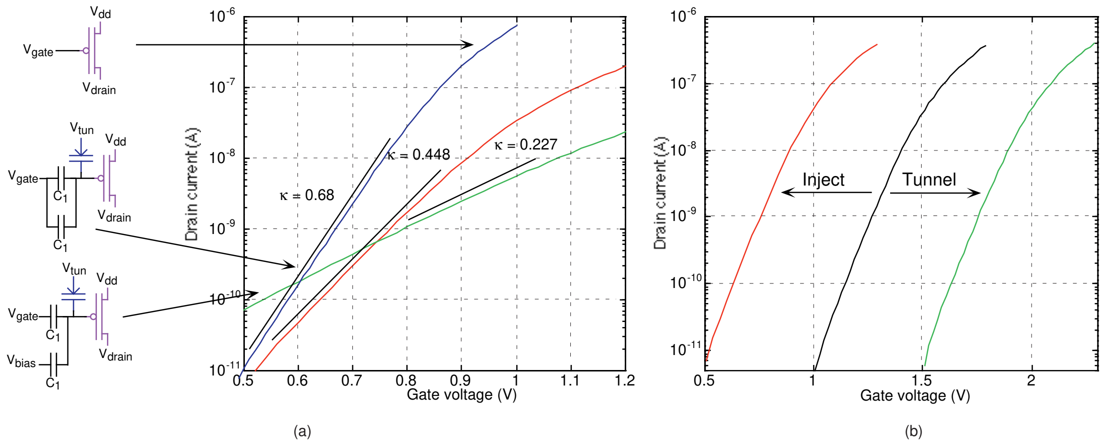


Figure 2: Illustration of the basic properties of floating-gate devices. (a) Comparison of the constant-charge behaviors. For all three circuits, we are sweeping the effective gate voltage between 0.5V and 1.2V below V_{dd}, and measuring the resulting source current. The first case is the baseline case of a measurement of a typical pFET in our 2.0μm process. In the other two cases, the capacitive coupling forms a capacitive voltage divider, resulting in a lower subthreshold slope. (b) The effect on the two capacitor curve before and after electron tunneling and hot-electron injection. Electron tunneling increases the device's effective threshold voltage, where hot-electron injection decreases the device's effective threshold voltage.

through the two floating-gate pFETs for

$$\text{Two capacitors}: I_s = I_{so} W \exp\left(-\frac{2C}{C_T}\frac{\kappa}{U_T}\Delta V_g\right), \ \text{One capacitor}: I_s = I_{so} W \exp\left(-\frac{C}{C_T}\frac{\kappa}{U_T}\Delta V_g\right). \tag{2}$$

where we define the weight, W, as the slow change in the source current due to changes in the floating-gate charge.

We 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. Alternately, we could use ultraviolet (UV) light to balance the charge on several floating gates, as used by other authors [18, 22]. Figure 2b shows the effect on the two capacitor curve before and after electron tunneling and hot-electron injection. Electron tunneling adds charge at the floating gate, as seen by the constant shift to higher gate voltages. 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. Hot-electron injection decreases charge at the floating gate, as seen by the constant shift to lower gate voltages. Injection occurs for large drain voltages; therefore, we can reduce the floating-gate charge during normal feedforward operation by raising the drain voltage.

1.1 Electron Tunneling Increases Floating-Gate Charge

Figure 3a shows the tunneling structure that we use to remove electrons from the floating-gate; electron tunneling produces positive floating-gate current. We tunnel electrons through a MOS capacitor rather than through a polysilicon-polysilicon capacitor because of the substantially better oxide quality, and therefore substantially improved reliability. The high quality of the gate oxide is the one constant for any CMOS process. Tunneling arises from the fact that an electron wavefunction has finite spatial extent. For a thin enough barrier, this spatial extent is sufficient for an electron to pass through the barrier. An electric field across the oxide, created by the voltage difference between the tunneling voltage, V_{tun}, and the floating-gate voltage, V_{fg}, reduces the thickness of the barrier to the electrons on the floating gate, and will allow some electrons to move through the oxide. The classic model of electron tunneling through a silicon–silicon-dioxide system

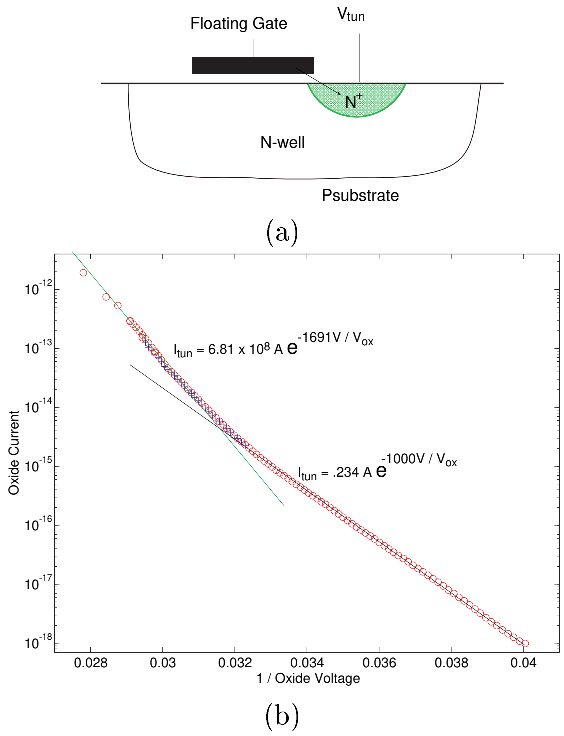


Figure 3: Tunneling in an nwell process. (a) The tunneling junction is the capacitor between the floating gate and the nwell (a MOScap); we use high-quality gate oxide to reduce the effects of electron trapping. Over a wide range of oxide voltage, most of the tunneling occurs between the floating gate and n^+ diffusion region, because the accumulated region at the corner of the floating gate is the source of higher electric fields. (b) Electron tunneling current versus 1/oxide voltage in a 2.0μm process with 42nm gate oxide. The two straight line fits are to the classic Fowler-Nordheim expression in (3). 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.

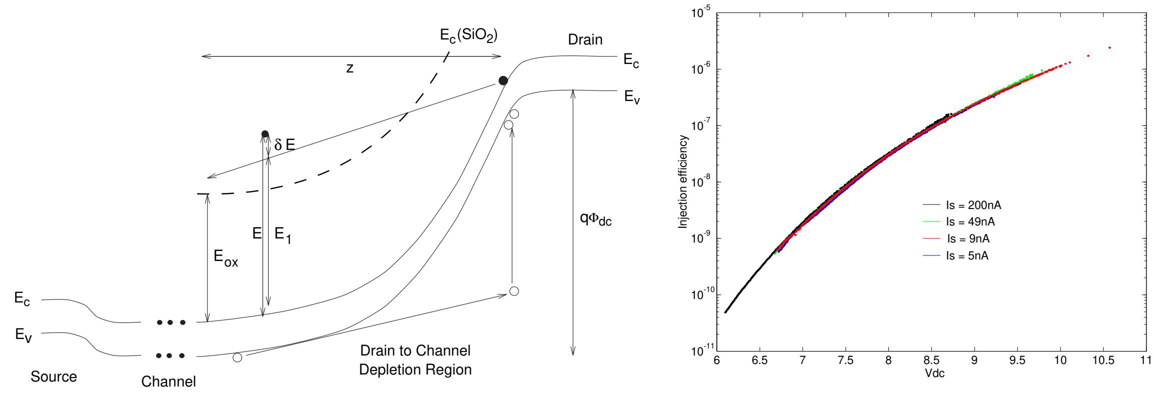
[24] results in an electron tunneling current given by

$$I_{tun} = I_0 \exp\left(-\frac{\mathcal{E}_o}{\mathcal{E}_{ox}}\right) = I_0 \exp\left(-\frac{t_{ox}\mathcal{E}_o}{V_{tun} - V_{fg}}\right), \tag{3}$$

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 [25]. Typical values for the oxide electric field range from 0.75V/nm to 1.0V/nm. Figure 3b shows a plot of tunneling current versus 1/oxide voltage ($V_{tun} - V_{fg}$) showing good agreement with (3) in two regions. The tunneling rate is modulated by both the input voltage and the charge on the floating gate, because both factors affect the floating-gate voltage. The two straight-line regions of the data in Figure 3b are due to differing tunneling processes that occur at different electric-field ranges. At low oxide voltage, electron tunneling is probably due to tunneling through intermediate traps [26], or due to tunneling through the junction edge. At high oxide voltages, electron tunneling is through the middle of the tunneling capacitor.

1.2 Hot-Electron Injection Decreases Floating-Gate Charge

One might wonder how pFETs, where the current carriers are holes, inject hot electrons onto the floating gate. Figure 4a 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. The hole impact-ionization current is proportional to the pFET source current, and is the exponential of a smooth function of the drain-to-channel potential (Φ_{dc}). 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, the electrons can be injected into the oxide and transported to the floating gate. The injection current is proportional to the hole impact-ionization current, and is the exponential of another smooth function of the voltage drop from channel to drain. Additional floating-gate dependences have little relevance to our present discussion, so we will neglect them [5].



(a) (b)

Figure 4: pFET hot-electron injection. (a) Band diagram of a subthreshold pFET transistor under conditions favorable for hot-electron injection. E_{ox} is the Si–SiO_2 barrier, which is 3.1eV for no field across the oxide. (b) Measured data of pFET injection efficiency versus the drain-to-channel voltage for four source currents. Injection efficiency is the ratio of injection current to source current. The injection efficiencies are nearly identical for the different source currents; therefore, they appear to be indistinguishable on the plot. At Φ_{dc} equal to 8.2V, the injection efficiency e-folds (increases by a factor of e) for a 250 mV increase in Φ_{dc}.

Figure 4b shows measured injection efficiency for four source currents; injection efficiency is the ratio of the injection current (I_{inj}) to source current (I_s). The measurements for four different source current values are nearly identical, which is consistent with injection efficiency being independent of source current. Injection efficiency is approximately an exponential of a linear function in Φ_{dc} over ranges spanning 1V to 2V. The slope of the curve on this exponential scale decreases with increasing Φ_{dc}. Using this linear approximation, we can model the hot-electron injection current for a changing gate and drain (ΔV_d) voltage as [5]

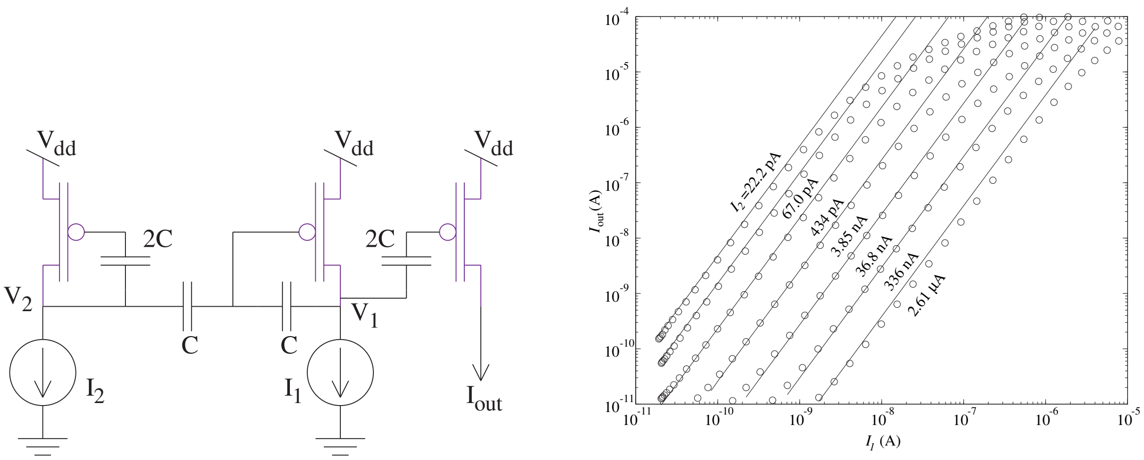
$$I_{inj} = I_{inj0} \left(\frac{I_s}{I_{s0}} \right)^{\alpha} \exp \left(\frac{-\Delta V_d}{V_{inj}} \right) = I_{inj0} \exp \left(-\frac{\alpha \kappa \Delta V_{fg}}{U_T} - \frac{\Delta V_d}{V_{inj}} \right), \tag{4}$$

where V_d is the drain voltage, and α is $1 - \frac{U_T}{V_{inj}}$. For a quiescent $\Phi_{dc} = 8.2$V, a typical value for V_{inj} is 250mV, and for α is 0.90. We have validated this model over several orders of magnitude in current and wide ranges in voltage elsewhere [8]. We previously derived an analytical model of the hot-electron transport in nFETs from Boltzmann transport that compares closely to experimental measurements of the impact-ionization and hot-electron-injection currents [6]; many of the results extend directly to a detailed model of pFET hot-electron injection.

2 Applications of Floating-Gate Devices and Circuits

2.1 Translinear Floating-Gate Devices

The floating-gate voltage, determined by charge stored on the floating-gate, can modulate a channel between a source and drain, and therefore can be used in computation. One can generate products of a power of input variables by exponentiating a weighted sum of the logarithm of the inputs. We have shown in earlier work that we could use these floating-gate devices to compute generalized translinear functions by a particular choice of capacitive couplings into the floating-gate device [10]; we will show the particular example shown in Fig. 5. Much of this work was inspired by the pioneering work of Shibata and Ohmi [17]. We use the fact that the subthreshold channel current is



(a) (b)

Figure 5: Translinear circuits built with floating-gate MOS devices [10] (a) Floating-gate MOS circuit that computes an output current, I_{out}, that is equal to I_1^2 / I_2. All capacitors are of equal size, and the tunneling junctions not drawn for clarity. (b) Measured data from this circuit with a curve fit to the ideal equation. [10]

proportional to the exponential of the floating-gate voltage and that we can use different capacitors to obtain different weighting functions.

We can derive the input-output relationship for the circuit in Fig. 5 from our basic pFET floating-gate equations. Assume that the devices are matched and that the charge on each floating-gate is identical, that is the weights are all identical. By directly applying the floating-gate equations, we get

$$\Delta V_2 = \frac{C_T U_T}{2C\kappa} \ln\left(\frac{I_2}{I_{so}W}\right), \tag{5}$$

$$\Delta V_1 = \frac{C_T U_T}{C\kappa} \ln\left(\frac{I_1}{I_{so}W}\right) - \Delta V_2 = \frac{C_T U_T}{2C\kappa} \ln\left(\frac{I_1^2}{I_2 I_{so}W}\right). \tag{6}$$

The circuit connecting V_1 and V_2 is an amplifier with capacitive feedback, as seen if I_1 is fixed in the above equation; we will discuss this further in Section 4. From these two equations, we get

$$I_{out} = I_{so}W \exp\left(\frac{2C\kappa \Delta V_2}{C_T U_T}\right) = I_{so}W \left(\frac{I_1^2}{I_2 I_{so}W}\right) = \frac{I_1^2}{I_2} \tag{7}$$

Figure 5b shows agreement between measured data and emperical curve fit to I_1^2 / I_2 over several orders of magnitude in current.

2.2 Floating-Gate Devices as memory elements

Figure 6 illustrates our nonvolatile electronic voltage biases. Modern analog, neuromorphic, or mixed-mode VLSI chips typically have large numbers of inputs and analog parameters, and the number of available pins is often a limiting factor in these systems. Often, these parameters and other biases are stored off chip as voltages programmed by potentiometers or other sources, one pin per variable. Pins are also needed for inputs and outputs. The number of pins available on a chip is limited by the perimeter of the chip, which grows as the square root of die area. By moving the analog parameters and circuit biases onto the chip, we save many pins for input, output, and diagnostics. If we think of potentiometers as easily modifiable voltages sources, then these programmable floating-gate voltage sources would be an electronic version of the pot, an e-pot,

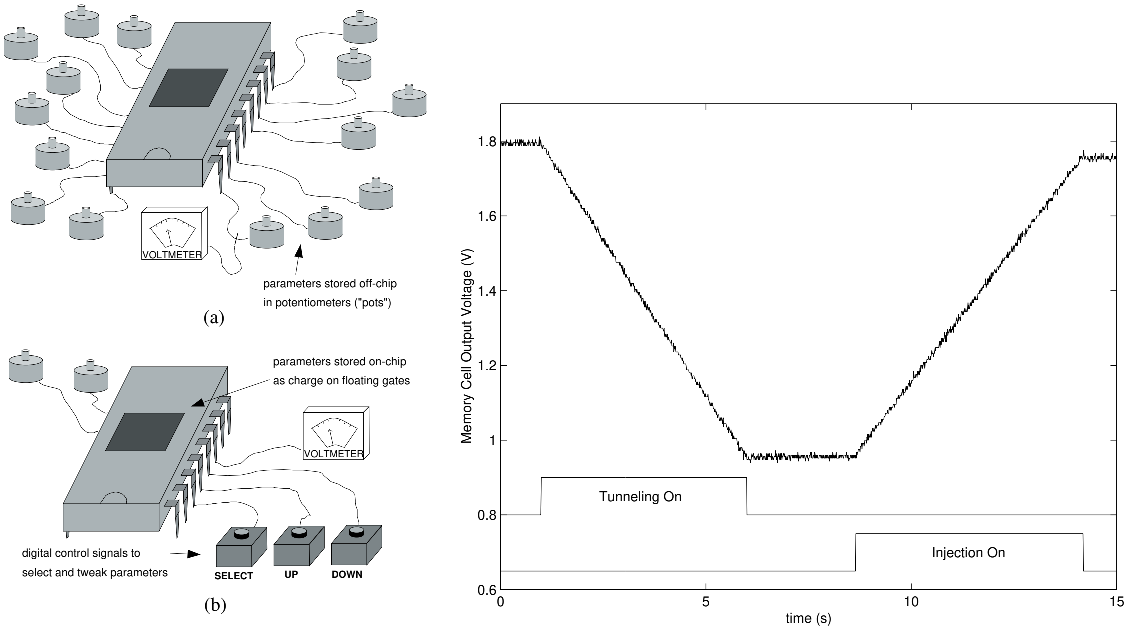
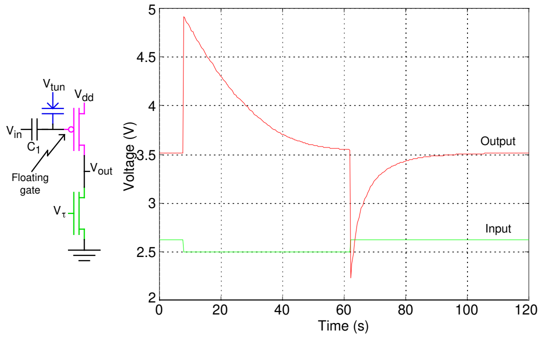


Figure 6: Electronic Potentiometers (e-pots) for non-volatile storage of voltage biases. (a) On typical analog VLSI chips, many of the pins are consumed by bias voltages set with off-chip potentiometers. (b) By storing these voltages on chip, many pins are freed for I/O. (c) Experimental measurement illustrating e-pot operation. Digital signals control tunneling and hot-electron injection, moving the e-pot output voltage up or down smoothly.

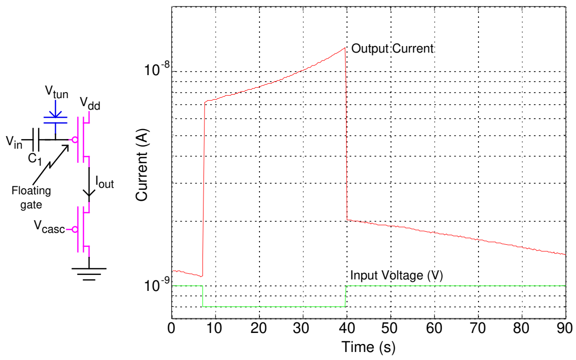
that can be placed on the chip itself. If large numbers of these cells can be placed on one chip, then they will supply more biases than could the few pins required to program them. Also, storing biases on chip could significantly reduce the resulting circuit board complexity.

We use floating-gate technology to eliminate off-chip-biasing voltages in the existing system by providing these voltages on chip with arrays of programmable floating-gate voltages. The first programmable analog memories stored arrays of analog quantities [20]; analog EEPROM chips have found applications in audio recorder chips [2]. We recently reported an array of floating-gate memory elements which can be individually programmed either up or down by straightforward controls [12], based upon our earlier floating-gate memory work [15]. Figure 6 shows experimental data from a single memory element where the voltage is decreased by inputing one digital signal to tunnel electrons off of a floating-gate memory and is increased by inputing a second digital signal to inject electrons onto the floating-gate memory. To select an individual memory element, we use high-voltage amplifiers to switch the high tunneling voltage to each tunneling junction, and another amplifier to switch a high drain voltage for hot-electron injection. These elements routinely achieved 13 bits of precision, primarily dependent upon the quality of the supporting analog circuitry, rather than our ability to control the transport of charge through the oxide. At a fundamental level, the number of voltage levels on the floating gate, where a voltage level is determined by the voltage that a single electron generates on the floating-gate capacitor, limits the precision.

One key issue in the programming of these values is to compensate for capacitive-coupling shifts due to changing the tunneling voltage through the tunneling capacitor. A related issue is the capacitive coupling of the drain voltage due to the gate-to-drain overlap capacitance. For this circuit, which is functional in 2.0μm and 1.2μm processes, we use an active capacitor compensation scheme to null out these capacitive effects.



(a)



(b)

Figure 7: (a) The pFET voltage-adapting circuit configuration. Response to an upgoing and a downgoing step input. The adaptation in response to an upward step results from electron tunneling; the adaptation in response to a downward step results from pFET hot-electron injection. (b) The pFET current-adapting circuit configuration. Response of the current-adapting floating-gate pFET circuit to an upgoing and a downgoing step input. This circuit configuration is unstable.

3 Adaptation in Floating-Gate Circuits

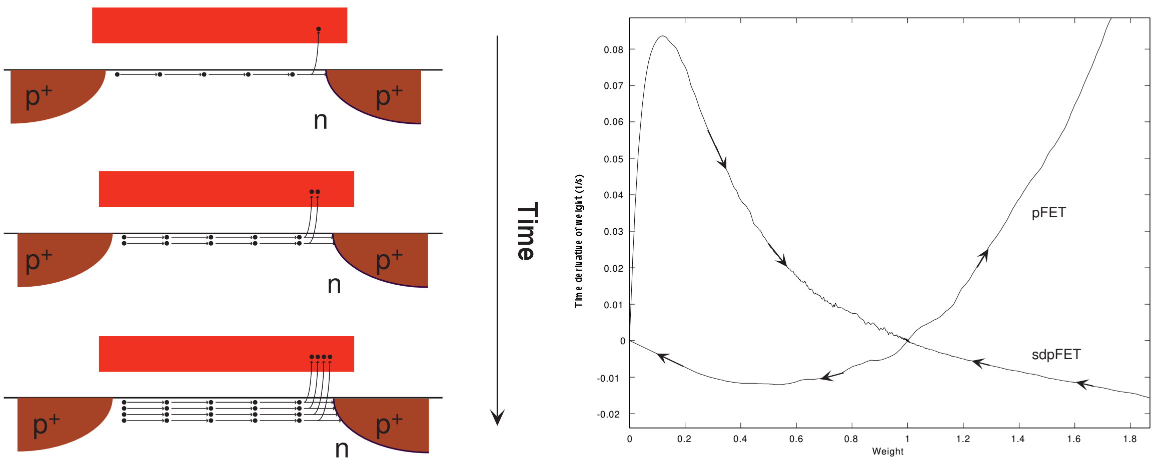
To this point, we have considered floating-gate devices only as programmable storage devices or as computing elements, but now we will discuss the greater win for floating-gate devices, that is that they can adapt their parameter as a function of the computations in which they participate. In this arena, a floating-gate device will be considered as a continuous-time circuit element computing at several timescales. The oxide currents can be arbitrarily smaller than the MOSFET's channel current; typically, the oxide currents are at least 10^4 times smaller than the channel currents. These small current levels give us slow programming rates, which are precisely what we need to build adaptive systems, because the adaptation should integrate over many presentations of the inputs, or over several periods of the input.

One major change is that we have not yet needed an accurate model of modeling the floating-gate charge, because a qualitative understanding of tunneling and injection modifying the floating-gate charge was sufficient. Recall that the weight, W is proportional to the source current, and proportional to the exponential of gate current. We can write down a differential equation modeling the weight by equating currents at the floating-gate, and taking the derivative versus time of the equation relating W and the floating-gate: [8]

$$\frac{U_T C_T}{\kappa_p I_{inj} W}\frac{dW}{dt} = -C_1\frac{dV_{in}}{dt} - C_2\frac{dV_d}{dt} - (W)^{1+\frac{U_T}{\kappa_p V_x}} + (W)^{1+\alpha}\exp\left(-\frac{\Delta V_d}{V_{in}}\right). \qquad (8)$$

Hot-electron injection adds electrons to the floating gate, thereby decreasing the weight, and electron tunneling removes electrons from the floating gate, thereby increasing the weight. Because we assume that the floating-gate pFETs have nonnegligible tunneling and injection currents, equilibrium is established when the tunneling current is balanced by the injection current. As a result, we assume that $I_{tun0} = I_{inj0}$ for the rest of our discussion.

We will first illustrate floating-gate adaptation using circuits employing a single floating-gate device. The dynamic behavior of a single floating-gate pFET can be characterized from that device's response in both a constant-current circuit and a constant-voltage circuit. Figure 7 shows two pFET circuits, one with the drain connected to a current source and one with the drain connected to a cascode transistor, as well as their respective responses to an upgoing and downgoing input voltage step. From (8), we can analyticaly predict the behaviors in Fig. 7. The constant-current (fixed W) pFET circuit is stable. For the pFET circuits in the constant-current configuration, the source



(a) (b)

Figure 8: (a) llustration of the feedback that hot-electron injection provides to a MOSFET's channel current. The pFET hot electron injection is a positive feedback mechanism. This process "strengthens" the transistor, and is a run-away process. (b) Plot of the time-derivative of W versus W for the pFET, and the source-degenerated floating-gate pFET. The arrows show the direction that the differential equation will take. This data shows that the floating-gate pFET will diverge from the W=1 steady state, but that source-degenerated floating-gate will stabilize to the W=1 steady state. The s-d pFET modifies the basic pFET dynamics by local negative-feedback.

current is fixed at I_{so} (W = 1), and the tunneling current is held constant at I_{tun0} because of the constant floating-gate voltage. Using these simplifications results in the model from (8) :

$$C_2 \frac{dV_{out}}{dt} = I_{tun0} \left(e^{-\Delta V_{out}/V_{inj}} - 1 \right). \tag{9}$$

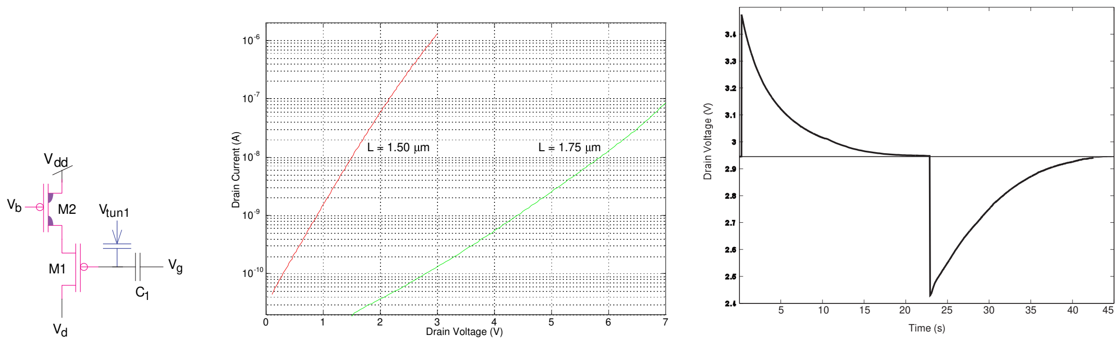
We showed previously that equations of this form can be solved analytically [5]. The trajectories of the pFET equation converge to the steady state at $\Delta V_{out} = 0$. The constant-voltage pFET circuit is unstable. For the pFET circuit in the constant-voltage configuration, the drain voltage of the pFET is fixed by the cascode transistor, and we simplify (8) to

$$\frac{C_T U_T}{\kappa_p I_{tun0}} \frac{dW}{dt} = -W^{1+\frac{U_T}{\kappa_p V_x}} + W^{1+\alpha} = W^{1+\frac{U_T}{\kappa_p V_x}} \left(W^{\alpha - \frac{U_T}{\kappa_p V_x}} - 1 \right). \tag{10}$$

The trajectories of the pFET differential equation converge toward $W = 0$, and diverge away from $W = 1$.

Hot-electron injection is an important adaptation mechanism, because floating-gate current is proportional to channel current and to an exponential function of drain voltage. Continuous-time adaptation occurs because the gate current is a function of the device parameters; therefore, floating-gate adaptation is a direct function of the computations being performed. In practical systems, we use a combination of electron tunneling and hot-electron injection to modify the floating-gate charge, but we encounter similar dynamics. Once the adaptation has converged, the floating-gate devices provide non-volatile storage of the final result.

Figure 8a shows electrons moving from the transistor channel to the floating gate by hot-electron injection. For a pFET device, injection means more electrons on the floating-gate, resulting in more current, and therefore resulting in more injection of electrons. This process results in ever increasing injection current. For floating-gate pFET devices, hot-electron injection is a positive feedback mechanism. We can graphically illustrate the floating-gate dynamics by plotting $\frac{dW}{dt}$ versus W, as shown in Fig. 8b. If $W < 1$, then we see that the derivative is negative, and therefore



(a) (b) (c)

Figure 9: The source-degenerated (s-d) floating-gate pFET. The s-d pFET modifies the basic pFET dynamics by local negative-feedback; the device converges to a stable equilibrium for either a constant current or a constant voltage. (a) Circuit diagram of the s-d pFET. The s-d pFET is comprised of a floating-gate pFET and a second ultra-short pFET, which provides feedback to the source terminal. We utilize the DIBL effect in short-channel MOSFETs to build a compact weakly exponential element. (b) Source current versus drain voltage for two short-channel pFETs in a 2μm process with the gate and source voltages at V_{dd}, and the drain voltage measured relative to V_{dd}. The DIBL results in an exponentially increasing current for subthreshold biases; for the 1.5μm pFET, the current increases an e-fold for a 276.1mV change in the drain voltage, and for the 1.75μm pFET, the current increases an e-fold for a 627.3mV change in the drain voltage. (c) The behavior of the current autozeroing circuit using a source-degenerated pFET. Unlike the floating-gate pFET, this circuit converges to its steady-state current.

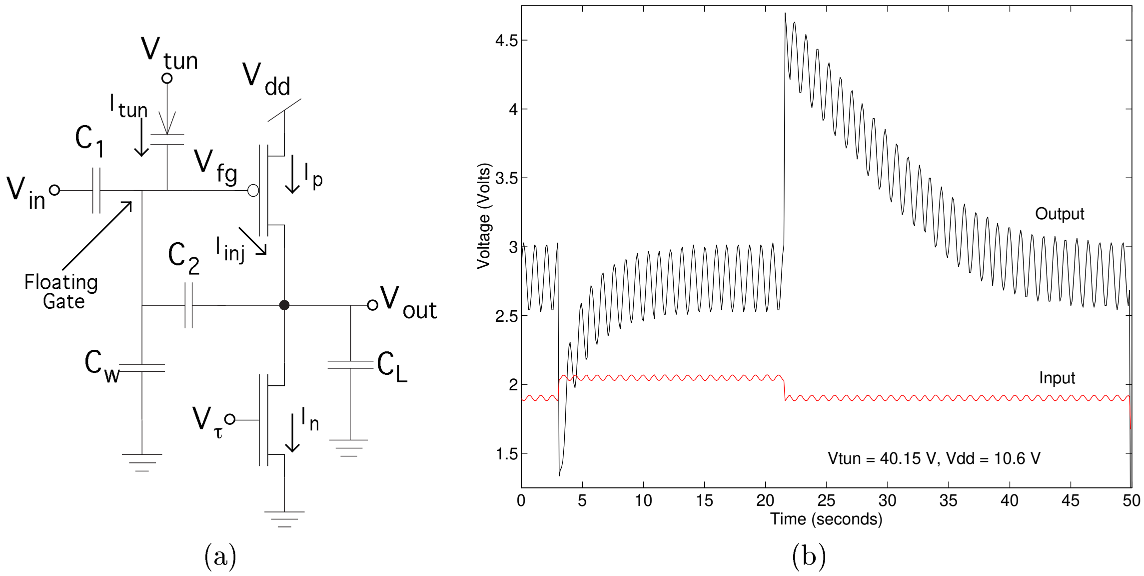
W decreases. If $W > 1$, then we see that the derivative is positive, and therefore W increases. Both statements are consistent with an unstable equilibrium due to positive feedback.

Figure 8b shows that in a source-degenerated (s-d) pFET device, hot-electron injection provides negative feedback in a self-limiting process. The s-d pFET device modifies the pFET device by adding local negative-feedback to the source terminal. The s-d pFET moves the source voltage to decrease the transistor's drain-to-source voltage by an amount large enough so that an increase in source current results in decreased hot-electon-injection current. Figure 9 shows the circuit model for the (s-d) floating-gate pFET [8]. The resulting feedback mechanisms due to the gate current dependencies of this s-d pFET make it a practical candidate for continuously interacting arrays of floating-gate circuits.

The s-d floating-gate pFET is a floating-gate pFET that uses a weak exponential element—the upper pFET in Fig. 9—to change the injection-current dependence upon channel current. We use a short-channel pFET with significant drain-induced barrier lowering (DIBL) to implement a device with a weak exponential dependence. A transistor that strongly exhibits DIBL shows an exponential change in current for a linear change in drain voltage, as we show in Fig. 9b. For an s-d floating-gate pFET, the constant-voltage and constant-current configurations are stable. From Fig. 8 we see that the s-d pFET tends to converge to its steady state at $W = 1$; if $W > 1$, the negative derivative decreases W towards 1, else if $W < 1$, the positive derivative increases W towards 1. Figure 9c shows measured drain current for an input voltage step; the drain current returns to its equilibrium level unlike the basic floating-gate pFET. The constant current configuration for the s-d floating-gate pFET is stable as is the basic pFET circuit [8].

4 Applications of Adaptation using a Single Floating-Gate Device

The autozeroing floating-gate amplifier (AFGA), shown in Fig. 10, is a simple example of an adaptive floating-gate pFET circuit. The AFGA uses tunneling and pFET hot-electron injection



(a) (b)

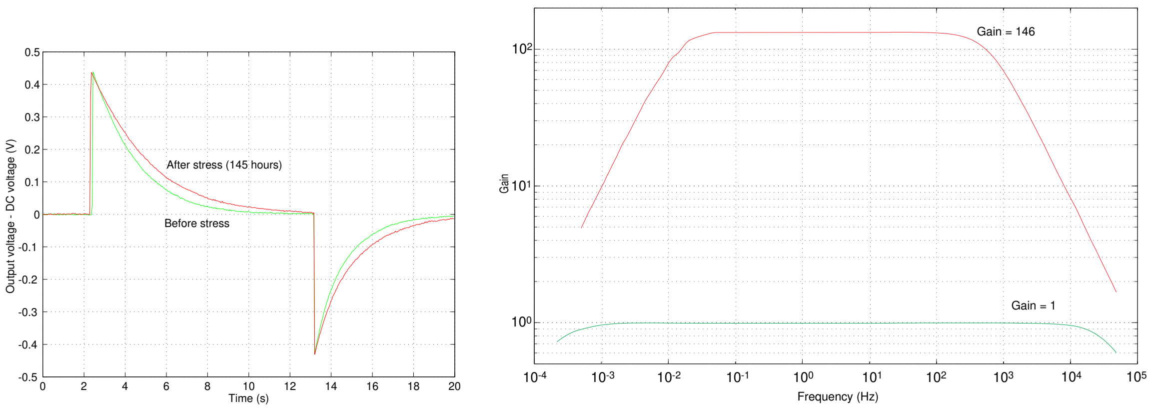
Figure 10: (a) Circuit diagram and measured data of an autozeroing floating-gate amplifier (AFGA) that uses pFET hot-electron injection. The ratio of C_2 to C_1 sets the gain of this inverting amplifier. Steady state occurs when the injection current is equal to the tunneling current. The capacitances C_w and C_L represent both the parasitic and the explicitly drawn capacitances. (b) Response of the AFGA to a 1Hz sinewave superimposed on a 19s voltage pulse. The AFGA has a closed-loop gain of 11.2, and a low-frequency cutoff at 100mHz. The 1Hz signal is amplified, but the much slower step is adapted away.

to adaptively set its DC operating point. The modulation of the pFET hot-electron injection by the output voltage provides the correct feedback to return the output voltage to the proper operating regime. The AFGA is based on the stability of the circuit configuration in Fig. 7a [5]. That amplifier has a gain of 11.2, and I_{tun0} is 50fA. Because of feedback applied to the floating gate, this adaptation is an inherent part of the circuit's operation—no additional control circuitry is required. The AFGA demonstrates how to use continuous-time, floating-gate adaptation in amplifier design, and is an example of how one of many classical engineering problems that is solvable using floating-gate techniques.

The physical properties of the tunneling and hot-electron injection mechanisms change with time. These processes are permanently modified as electrons pass through the oxide, creating electron traps. We investigated the long-term changes by performing an accelerated stress experiment, where we operated an AFGA continuously for 145 hours with an average τ_l of 1.7s. When an AFGA is used as an amplifier or as a low-pass filter, a more reasonable τ_l would be at least several minutes; therefore, this experiment represents the stress of operating the AFGA continuously for a few years. Figure 11a shows the square-wave response of the AFGA before and after this lifetime test. The adaptation time constant has increased noticeably, but the general behavior is unaffected. Most of the long-term change is due to changes in the tunneling junction, which is probably a consequence of electron trapping.

The AFGA transfer function is bandpass, with the low-frequency cutoff set by the equilibrium tunneling and injection currents, and the high-pass cutoff independently set by the equilibrium pFET and nFET channel currents. For timescales between the adaptation and integrating regimes, the AFGA acts as an amplifier. Figure 11b shows the measured AFGA frequency responses. In the adaptation regime, the AFGA behaves as a high-pass filter:

$$\frac{V_{out}(s)}{V_{in}(s)} = -\frac{C_1}{C_2}\frac{s\tau_l}{1+s\tau_l}, \qquad \tau_l = \frac{C_2 V_{inj}}{I_{tun0}}. \tag{11}$$



(a) (b)

Figure 11: The effect of long-duration AFGA operation. (a) The responses to an upgoing and downgoing voltage step before and after 145 hours of operation. We plot the difference in the output voltage from the equilibrium DC level as a function of time; the equilibrium output voltage increased slightly over the 145 hours of operation. Most of the long-term change comes from the tunneling junction, which is probably due to oxide trapping. The effect of an input signal only slightly modifies the results of this experiment. (b) Frequency response for two AFGAs with different gains. For both the high- and low-gain AFGA, $C_1 + C_2$ is approximately constant. For the high-gain AFGA, τ_l is 20mHz, and τ_h is 600Hz; for the low-gain AFGA, τ_l is 300μHz and τ_h is 40kHz.

Because the gate currents are small, the circuit exhibits a high-pass characteristic with a cutoff frequency less than 1 Hz. In the integrating regime, the AFGA behaves as a low-pass filter:

$$\frac{V_{out}}{V_{in}} = -\frac{C_1}{C_2}\frac{1 - A_h \tau_h s}{1 + \tau_h s}, \qquad \tau_h = \frac{\left(C_T C_o\right) - C_2^2\right) U_T}{\kappa C_2 I_\tau}, \tag{12}$$

where $C_T = C_1 + C_2 + C_w$, and C_o is the total capacitance connected to the output node ($C_o = C_2 + C_L$). This transfer function includes the effects of parasitic and load capacitances. 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.

We 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 [5]. Careful choices of capacitors allow the one to design for particular linear range, noise, and dynamic range specifications. By increasing C_w, we can reduce the change in the floating-gate voltage, thereby increasing the amplifier's input and output linear range. 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, we can reduce significantly the low-frequency noise generated in the AFGA. We previously calculated the AFGA's dynamic range, the ratio of the maximum possible linear output swing to the total output-noise power, as [5]

$$\text{Dynamic Range} \; = \; \frac{V_{Lo}^2}{2\hat{V}_{out}^2} \approx \frac{\kappa}{2q} V_{Lo} C_o. \tag{13}$$

The dynamic range varies inversely with C_2; therefore a high-gain amplifier will have a larger dynamic range than will the low-gain amplifier for the same values of C_1, C_w, and C_L.

We have used these AFGA techniques to sense changes in currents, resistance, or capacitance. We used floating-gate pFETs to elegently extended the winner-take-all circuit to be short-term sensitive due to relative changes in inputs. Another previous application is using an AFGA as a front-end for an olfactory sensor [13]. The autozeroing second-order section (AUTOSOS) is one of several applications that use these AFGAs. We can electronically change the frequency response in an AutoSOS to get resonance response at either high or low frequencies [9]. We have used floating-gate second-order sections to build cochlear models [23].

5 Adaptation in Multiple Floating-Gate Devices

Figure 12 shows two circuits each with multiple interacting floating-gate pFETs that are continuously adapting. In both circuits, the pFETs share the same drain voltage, and the total current through all pFETs is fixed through a bias current set by an nFET current source. Several other feedback configurations can be found elsewhere [8].

In both cases, the output voltage (V_{out}) quickly returns to equilibrium, because the pFET with the largest current drives the system, making that pFET behave like an AFGA. Once the output voltage reaches equilibrium, it stays at the same value when the inputs are constant; therefore, each device will act as an independant single-device circuit with a fixed drain voltage.

Figure 12a shows that the two pFETs compete for the available bias current; if the two starting currents are equal, then, over time, the current in one pFET will decrease to 0, and the current in the other pFET will increase towards the bias current. If the current of the losing pFET is brought slightly above the current of the winning pFET, what was the losing pFET will adapt to being the winning pFET. A coupled network of floating-gate pFET will have at most one non-zero channel current for constant inputs.

Figure 12 shows experimental measurements of four s-d pFET transistors coupled through their drain terminals. We can measure the channel currents indirectly by measuring the source voltages (V_1, V_2, V_3, and V_4), because the current through the DIBL pFETs is a known exponential function of these source voltages. We then see that after applying a voltage step to all four inputs, all channel currents return to their equilibrium values. The output voltage also settles to its steady-state level. The individual source currents have settled to non-zero steady-states, which would only occur if the gate-current feedback was stabilizing in all changes in the gate voltage and in the drain voltage. The four s-d pFET devices (synapses) cooperate for the entire available channel current; even if the starting weights are set far apart, then, over time, both weights will converge to their equilibrium levels, with each transistor receiving a finite amount of the bias current.

6 Acknowledgements

The authors also wish to acknowledge C.Mead for all of his inspiration, guidance, and support through much of this work. The authors also wish to acknowledge all the insights and collaboration of our colleagues: A. Apsel, J. Dugger, R. Harrison, and T. Stanford.

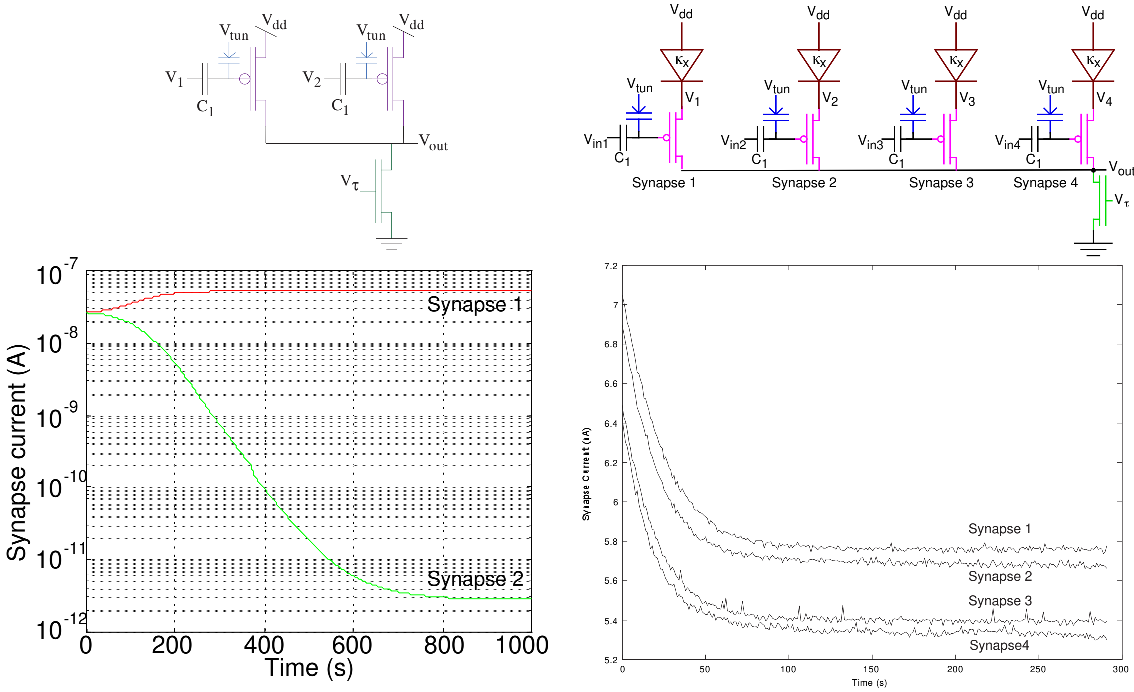
References

[1] D. Kahng and S.M. Sze, "A floating-gate and its application to memory devices," *The Bell System Technical Journal*, vol. 46, no. 4, 1967, pp. 1288-1295.

[2] *ISD Data Book: Voice Recording and Playback ICs*, ISD, San Jose, 1996, 2nd edition.

[3] P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single Transistor Learning Synapses," *Advances in Neural Information Processing Systems 7*, MIT Press, 1995, pp. 817-824. Also at *http://users.ece.gatech.edu/~phasler*.

[4] P. Hasler, B.A. Minch, C. Diorio, and C. Mead, "An autozeroing amplifier using pFET hot-electron injection," *Proceedings of the International Symposium on Circuits and Systems*, Atlanta, vol.3, 1996, pp. 325-328. Also at *http://users.ece.gatech.edu/~phasler*.



(a) (b)

Figure 12: (a) Circuit and experimental data of two continuously adapting pFET floating-gate for fixed inputs that are coupled at the drain with a current source. The capacitances are equal for this structure. This experiment starts the channel currents near each other with fixed inputs, and V_τ setting a bias current. Even though the synapse currents initially start near each other, I_1 wins and I_2 loses. I_2 decreases as a linear exponential in time due to the constant tunneling current at the floating gate. The measured I_2 saturates due to the surrounding leakage currents; the floating-gate voltage continues to increase with time. (b) Circuit diagram and experimental measurements of the four-input source-degenerated pFET synapse. This figure shows that the stabilizing behavior for two synapses is extendible to multiple synapses. Experimental measurements show the s-d pFET drain-current responses due to an input step applied to all four synapses.

[5] P. Hasler, B.A. Minch, C. Diorio, and C.A. Mead, "An autozeroing floating-gate amplifier," *Accepted for publication in Circuits and Systems II: Digital and Analog Signal Processing* Also will be at *http://users.ece.gatech.edu/∼phasler*.

[6] P. Hasler, A. Andreou, C. Diorio, B. A. Minch, and C.A. Mead, "Impact ionization and hot-electron injection derived consistently from Boltzmann transport," *VLSI Design*, in Press.

[7] P. Hasler, C. Diorio, and B. A. Minch, "A four-quadrant floating-gate synapse," *Proceedings of IEEE International Symposium on Circuits and Systems*, Monterey, 1998. Also at *http://users.ece.gatech.edu/∼phasler.*

[8] P. Hasler, Foundations of Learning in Analog VLSI, California Institute of Technology, February 1997. Also at *http://users.ece.gatech.edu/∼phasler.*

[9] P. Hasler, T. Stanford, B. A. Minch, and C. Diorio, "An Autozeroing Floating-Gate Second-Order Section," *Proceedings of IEEE International Symposium on Circuits and Systems*, Monterey, 1998. Also at *http://users.ece.gatech.edu/∼phasler.*

[10] B. A. Minch, C. Diorio, P. Hasler, and C.A. Mead, "Translinear circuits using subthreshold floating-gate MOS transistors," *Analog Integrated Circuits and Signal Processing*, vol. 9, no. 2, 1996, pp 167-179. Also at *http://users.ece.gatech.edu/∼phasler.*

[11] B.A. Minch, C. Diorio, P. Hasler, and C.A. Mead, "The matching of small capacitors for analog VLSI," *Proceedings of the International Symposium on Circuits and Systems*, Atlanta, vol. 1, 1996, pp. 238-241. Also at *http://users.ece.gatech.edu/~phasler.*

[12] R.R. Harison, P. Hasler, and B.A. Minch, "Floating-gate CMOS analog memory cell array," *Proceedings of IEEE International Symposium on Circuits and Systems*, Monterey, 1998. Also at *http://users.ece.gatech.edu/~phasler.*

[13] A. Apsel, T. Stanford, and P. Hasler, "An adaptive floating-gate olfaction sensor," *Proceedings of IEEE International Symposium on Circuits and Systems*, Monterey, 1998. Also at *http://users.ece.gatech.edu/~phasler.*

[14] C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A complementary pair of four-terminal silicon synapses," *Analog Integrated Circuits and Signal Processing*, 1997.

[15] C. Diorio, S. Mahajan, P. Hasler, B. A. Minch, and C. Mead, "A high resolution non-volatile analog memory cell," *Proceedings of the International Conference of Circuits and Systems*, Seattle, 1995, pp. 2233-2236.

[16] A. Thomsen and M.A. Brooke, "A floating gate MOSFET with tunneling injector fabricated using a standard double-polysilicon CMOS process," *IEEE Electron Device Letters*, vol. 12, 1991, pp. 111-113.

[17] T. Shibata and T. Ohmi, "A functional MOS transistor featuring gate-level weighted sum and threshold operations," *IEEE Transactions on Electron Devices*, vol. 39, no. 6, 1992, pp. 1444-1455.

[18] R.G. Benson and D.A. Kerns, "UV-activated conductances allow for multiple time-scale learning," *IEEE Transactions on Neural Networks*, vol. 4, no. 3, 1993, pp. 434-440.

[19] K. Yang and A.G. Andreou, "A Multiple Input Differential Amplifier Based on Charge Sharing on a Floating-Gate MOSFET," *Journal of Analog Integrated Circuits and Signal Processing*, vol. 6, no. 3, 1994, pp 197-208.

[20] H.V. Tran, T. Blyth, D. Sowards, L. Engh, B.S. Nataraj, T. Dunne, H. Wong, V. Serin, T. Lam, H. Hazarian, and G. Hu, "A 2.5V 256-level non-volatile analog storage device using EEPROM technology," *Proceedings of IEEE International Solid-State Circuits Conference*, 1996, pp. 270-271.

[21] J. Ramirez-Angulo, "+-0.75V BiCMOS Four Quadrant Analog Multiplier with Rail-Rail Input Signal-Swing," in *Proceedings of the 1996 IEEE International Symposium on Circuits and Systems*, Atlanta, GA, vol. 1, 1996, pp. 242-245.

[22] Y. Berger, D.T. Wisland, T.S. Lande, and S. Mikkelsen, "Ultra low-voltage digital floating-gate UVMOS Circuits," *Proceedings of IEEE International Symposium on Circuits and Systems*, Monterey, 1998.

[23] C.A. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, MA, 1989.

[24] M. Lenzlinger and E.H. Snow, "Fowler–Nordheim tunneling into thermally grown $\mathrm{Si0_2}$," *Journal of Applied Physics*, vol. 40, no. 1, 1969, pp.278–283.

[25] C.A.Mead, "Scaling of MOS technology to submicrometer feature sizes," *Journal of VLSI Signal Processing,* vol. 8, 1994, pp. 9-25.

[26] Y. Xu, *Electron Transport through Thin Film Amorphous Silicon — A Tunneling Study*, Ph.D. dissertation, Stanford University, 1992.

[27] T.A. Fjeldly and M. Shur, "Threshold voltage modeling and the subthreshold regime of operation of short-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 40, no. 1, 1993, pp. 137-145.