A Precision CMOS Amplifier Using Floating-Gate Transistors for Offset Cancellation

Venkatesh Srinivasan, *Member, IEEE*, Guillermo J. Serrano, Jordan Gray, and Paul Hasler, *Senior Member, IEEE*

***Abstract*—A long-term offset cancellation scheme that enables continuous-time amplifier operation is described. Offset cancellation is achieved by programming floating-gate transistors that form an integral part of the amplifier's architecture. The offset voltage of a single-stage folded cascode amplifier has been programmed to a minimum of ± 25 μV in a 0.5 μm digital CMOS process. The long-term offset voltage drift has been calculated to be less than 0.5 μV over a period of 10 years at 55 °C from a thermionic emission model for floating-gate charge loss. The offset voltage varies by a maximum of 130 μV over a temperature range of 170 °C, thereby making this a viable approach to offset cancellation.**

***Index Terms*—Charge retention, floating-gate drift, floating-gate transistors, input offset voltage, offset cancellation.**

I. Offset Removal Techniques

MISMATCHES between MOS transistors pose a serious challenge to analog circuit designers and most commonly manifest themselves as an offset voltage in operational amplifiers. Techniques commonly used to reduce the offset voltage include auto-zeroing, correlated double sampling and chopper stabilization [1]. Auto-zeroing and correlated double sampling are techniques applicable to sampled data systems while chopper stabilization allows continuous-time operation of the amplifier. Resistor trimming using laser trims is another popular approach. This, however, is usually expensive. Another technique includes using current-mode digital-to-analog converters (DACs) to compensate for amplifier offsets by adjusting amplifier load currents [2].

In this paper, a floating-gate based offset cancellation scheme is presented that results in a continuous-time operation of the amplifier with long-term offset cancellation that obviates the need for any refresh circuitry. A prototype amplifier has been fabricated with its offset voltage reduced to 25 μV. The use of floating-gate transistors for correcting mismatches in analog circuitry is particularly advantageous as it offers programmability, long-term retention and can be fabricated in a standard digital CMOS process. This approach involves no sampling and hence avoids such issues as charge injection, clock feedthrough and undersampled wideband noise that are serious limitations to autozeroing and correlated double sampling [1], [3]. Also, unlike chopper stabilization [1], the proposed scheme is not limited to low-bandwidth applications, while at the same time offering continuous-time operation with comparable offset reduction.

Manuscript received December 8, 2005; revised October 12, 2006.

V. Srinivasan is with Texas Instruments Incorporated, Dallas, TX 75243 USA (e-mail: v-srinivasan@ti.com).

G. Serrano, J. Gray, and P. Hasler are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250 USA (e-mail: gserrano@ece.gatech.edu; jgray@ece.gatech.edu; phasler@ece.gatech.edu).

Digital Object Identifier 10.1109/JSSC.2006.889365

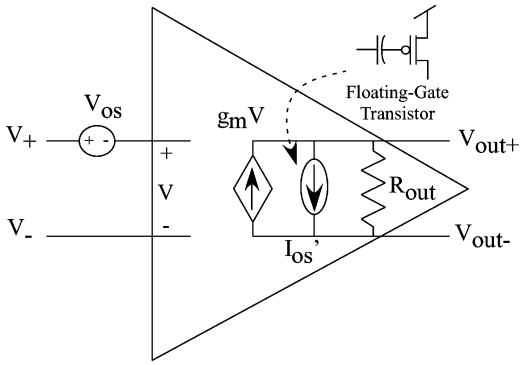

Fig. 1. Offset cancellation macromodel. The offset voltage of the amplifier V_{os} is cancelled by programming an offset current $I_{os'}$ in the opposite direction on floating-gate transistors.

The proposed scheme involves using floating-gate transistors as both an integral part of the circuit of interest and as programmable elements. Fig. 1 shows a conceptual representation of the proposed scheme applied towards offset cancellation in an operational amplifier. Floating-gate transistors are used as programmable current sources $(I_{os'})$ that provide offset compensation while being a part of the amplifier of interest during normal operation. Such an approach results in a compact architecture with a simple design strategy that avoids the overhead of using floating-gate transistors as separate trimming elements as in [4] and [5] or current-mode DACs as trimming elements [2]. Also, the proposed offset cancellation scheme is independent of other amplifier parameters, unlike other approaches [1], [6] and the offset cancellation by itself dissipates no additional power.

Section II describes floating-gate transistors and the programming scheme used to modify charge on the same. Sections III and IV discuss two key aspects of a floating-gate device, namely, the programming precision and charge retention. The use of floating-gate transistors as part of a folded-cascode amplifier is described in Section V along with a theoretical analysis of the input referred offset voltage and its temperature dependence. Section VI presents experimental results for the proposed amplifier fabricated in a 0.5 μm standard digital CMOS process. Finally, Section VII compares the proposed scheme with alternate approaches and Section VIII concludes by summarizing the results and the implications of the proposed approach.

II. Floating-Gate MOS Transistor

Floating-gate transistors are commonly used as nonvolatile memory elements in EEPROMs [7], [8]. A floating-gate MOS

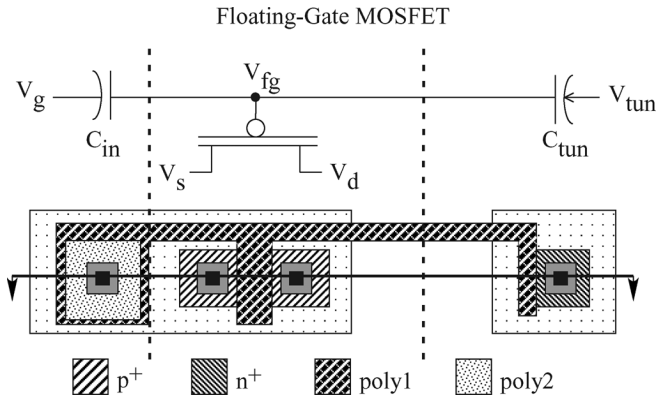

Fig. 2. Circuit schematic and layout of a pFET floating-gate transistor. The floating-gate node V_{fg} is completely surrounded by SiO_2 and external inputs are coupled onto the floating node through an input capacitor $C_{\rm in}$. The capacitor $C_{\rm tun}$ is used for Fowler–Nordheim tunneling. The input capacitor is implemented using a poly–poly capacitor, while the tunneling capacitor is implemented using a MOS capacitor.

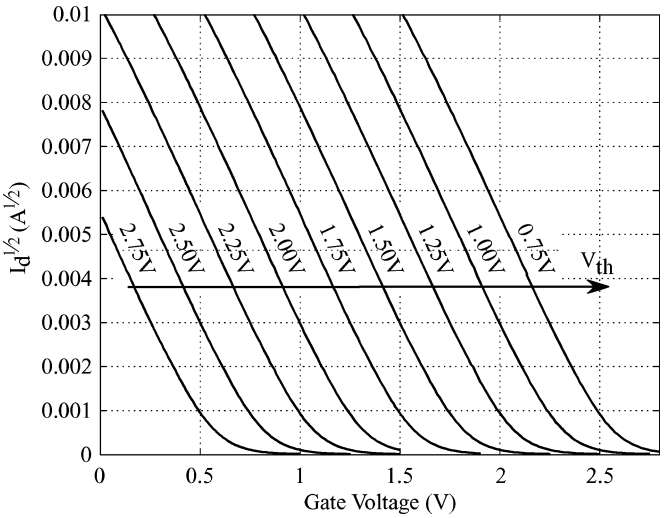

Fig. 3. Programming a floating-gate pFET transistor. A floating-gate pFET has been programmed using a combination of hot-electron injection and Fowler–Nordheim tunneling. Hot-electron injection adds electrons to the floating gate, thereby decreasing the threshold voltage of the device. Fowler–Nordheim tunneling removes electrons from the floating gate and leads to an increase in the threshold voltage.

transistor is a transistor whose polysilicon gate is completely surrounded by SiO_2, a high-quality insulator. This creates a potential barrier that prevents charge stored on the floating gate from leaking from the floating node. Fig. 2 shows the circuit schematic and layout of a single-poly floating-gate pMOS transistor. In order to maintain the nonvolatile charge storage of the floating gate, external inputs are capacitively coupled through an input capacitor $C_{\rm in}$. It should be noted that the second polysilicon layer shown in Fig. 2 is used primarily to implement the input capacitor. The tunneling capacitor $C_{\rm tun}$ is implemented using the gate oxide between the gate polysilicon and n-well.

A. Programming a Floating-Gate Transistor

Programming a floating-gate transistor involves adding or removing charge from the floating gate, thereby modulating the threshold voltage of the device. This is achieved through the physical phenomena of hot-electron injection that adds electrons to the floating gate and Fowler–Nordheim tunneling [9] that is used to remove electrons. Using hot-electron injection and tunneling, a floating-gate pFET transistor has been programmed to different threshold voltages with their magnitudes ranging from 0.75–2.75 V as demonstrated in Fig. 3. It should be noted that the absolute value of the threshold voltage of a pFET device that is not a floating gate in the 0.5 μm process used is 0.9 V. Fig. 3 demonstrates the wide range in programming capabilities of the floating-gate device.

The logarithmic nature of tunneling makes precision programming time-consuming. Techniques have been proposed to improve the speed and precision of tunneling-based programming [10], [11]. However, in this work, tunneling is used primarily as a global erase and precision programming is achieved through hot-electron injection. Such a scheme has a number of advantages over a tunneling-based programming scheme as in [4] and [5]. These include avoiding special processing steps such as ultrathin tunneling oxides and high voltages of both positive and negative polarities.

Hot-electron injection occurs in pFETs when carriers are accelerated to a high enough energy level to surmount the Si-SiO_2 barrier. At high drain currents, electrons are created at the drain edge of the drain-to-channel depletion region via hot-hole impact ionization. These electrons travel back into the channel region, gain sufficient kinetic energy such that they cross the Si-SiO_2 barrier and are injected onto the floating gate [12]. Conditions conducive for injection are created when the transistor experiences a high source–drain potential and when there is channel current flowing through the device.

Fig. 4(a) shows the use of floating-gate transistors as part of analog circuitry. During normal operation, a digital *Low* is applied to *prog*, thereby switching the floating-gate transistors into the circuit of interest. The operating V_{DD} is 3.3 V during normal operation. Programming is achieved by first isolating the floating-gate transistor from the rest of the circuitry such that one can access the gate and drain terminals of the device. This is achieved by applying a digital *High* to the *prog* terminal. The drain of the floating-gate transistor of interest that needs to be programmed is then switched to the external drain terminal V_d using the digital selection circuitry shown. The drains of the unselected devices are tied to V_{DD}. It should be noted that all floating-gate transistors share the same gate terminal during program mode. The tunneling terminal is shared among all floating-gate devices as well.

Fig. 4(b) shows a pictorial representation of the hot-electron injection process. To perform hot-electron injection on a floating-gate transistor, the chip is ramped up such that V_{DD} is increased to 6.5 V with all other voltages increased with respect to V_{DD} as well. Next, the high fields necessary for injection are created by pulsing down the drain voltage (V_D) for a certain amount of time $t_{\rm pulse}$ such that a high source–drain voltage appears across the device. Typical V_{SD} voltages used for hot-electron injection range from 4–6.5 V for a 0.5 μm CMOS process. After injection is completed, the chip is ramped down such that all voltages are restored to their original values. The number of electrons injected and hence the change in the drain current is a

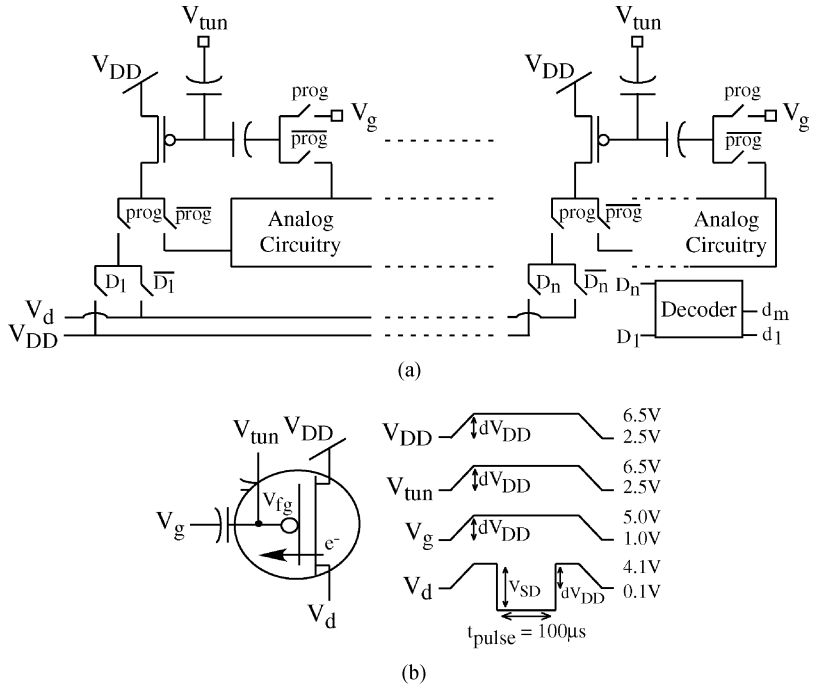

Fig. 4. Floating-gate transistors in analog circuitry and hot-electron injection. (a) The use of multiple floating-gate transistors as part of analog circuitry is shown. Applying a digital *High* to *prog* switches the floating-gate transistors into program mode. The floating-gate transistor of interest is then selected using the digital selection circuitry. (b) Pictorial representation of the hot-electron injection process.

function of the source–drain potential V_{SD} and the time interval $t_{\rm pulse}$ for which V_{SD} is held high enough.

Automatic programming of floating-gate transistors is achieved using a programming algorithm that calculates the value of the source–drain potential and the number of pulses required (with $t_{\rm pulse}$ held fixed) such that a target current is reached without any overshoot. The value of V_{SD} for a given pulse interval is estimated from the relationship between the initial drain current and the desired target current. This is obtained from a first-principles model for hot-electron injection and is described in detail in [12] and [13]. A single initial calibration step that characterizes the hot-electron injection rates for a given process is performed such that the algorithm can predict an optimal value of V_{SD} during injection. Programming proceeds by first measuring the initial drain current of the device. This is used by the programming algorithm along with the target drain current to calculate the optimal value of V_{SD} for a fixed pulse interval of about 100 μs. The chip is then ramped up and the calculated V_{SD} is applied. The chip is then ramped down and the drain current is measured again. If the measured current is different from the target current, the algorithm calculates a new V_{SD} value and injection is performed again. The above steps are repeated until the drain current of the device reaches the target value within a pre-defined tolerance. Since the injection rates vary between devices, programming is performed asymptotically such that overshoot is avoided. Typical convergence to target to within a 0.1% accuracy takes about 7–12 pulses.

Fig. 5 [13] demonstrates the programming algorithm when the drain current of the device is programmed to different target currents (logarithmical spaced) from an initial current of 10 nA. As can be observed, typically around 7 pulses are required to reach the target with the worst case of 11 pulses for a programming change of two orders of magnitude. A pulsewidth of 100 μs has been used in this case.

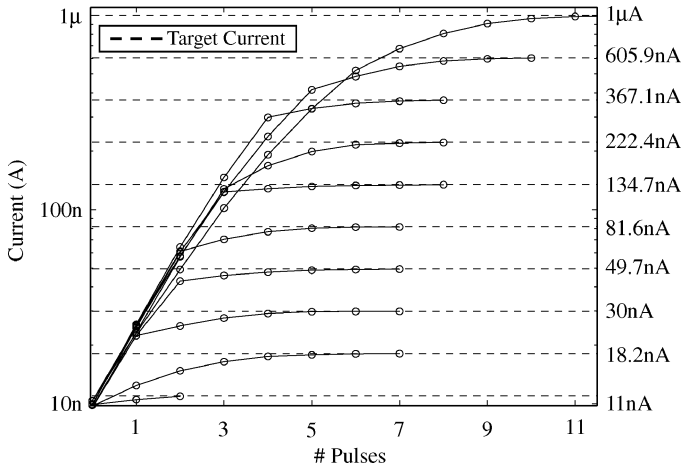

Fig. 5. Programming algorithm [13]. The convergence of the programming algorithm for different target currents from an initial starting current of 10 nA is shown. The algorithm convergence to within 0.1% of the target current in all cases with the pulsewidth being 100 μs.

III. Floating-Gate Programming Precision

The accuracy to which one can program floating-gate transistors to a target current depends on the smallest drain current change that can be programmed onto a floating-gate device. In order to estimate the design choices available to improve programming precision, a figure of merit (FOM) is defined as

$$\text{FOM} = -\log_2\left(\frac{\Delta I}{I}\right) \tag{1}$$

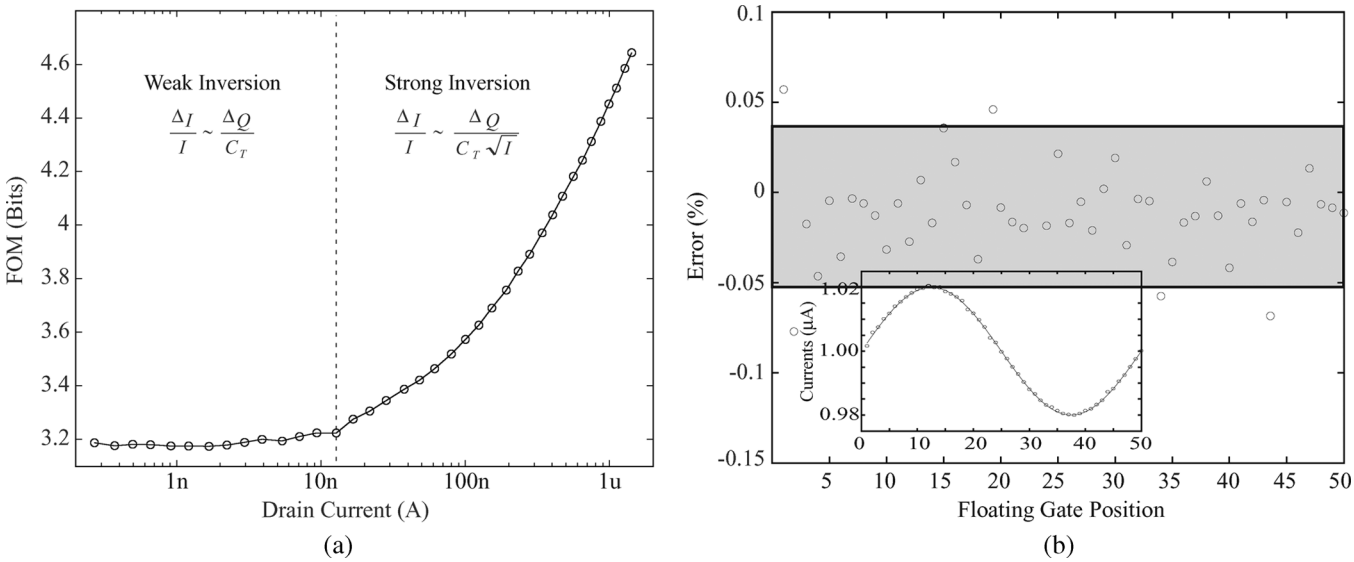

Fig. 6. Programming precision. (a) The FOM is independent of the drain current in the weak inversion region of operation and increases as the transistor enters the strong inversion regime. The experimental results are consistent with the theoretical prediction. (b) Programming a 20 nA sinusoid riding on a DC value of 1 μA is shown along with the percentage error between the programmed current and the desired target. As can be observed, an error of $\pm 0.05\%$ has been achieved.

where ΔI is the minimum programmable change in drain current that is necessary to meet a system level accuracy specification and I is the bias current of the floating-gate transistor. It should be noted that such a definition results in the FOM being represented in the familiar binary system, as the number of bits of accuracy achievable. In the discussion below, the FOM is related to floating-gate circuit parameters for operation in both the weak and strong inversion regimes such that the floating-gate transistor can be designed to achieve the required bits of precision.

Consider a floating-gate pFET operating in the weak inversion regime. The source–drain current of the device, ignoring Early effect, is given by [14]–[16]

$$I = I_o' \exp\left(\frac{-\kappa\,(V_{fg} - |V_{To}|)}{U_T}\right)\exp\left(\frac{V_s}{U_T}\right)\exp\left(\frac{-(1-\kappa)V_b}{U_T}\right) \tag{2}$$

where I_o' is the pre-exponential constant, U_T is the thermal voltage given by kT/q, V_{fg}, V_s and V_b are the floating-gate, source, and bulk voltages referenced to ground, respectively, $\kappa = C_{ox}/(C_{ox} + C_{dep})$ and V_{To} is the threshold voltage of the device referenced to the bulk terminal. Detailed analytical expressions for the various terms in the above equation, such as I_o' and V_{To}, can be found in [15] and [16].

Now, for a ΔV_{fg} change in the floating-gate voltage due to programming, a ΔI change in drain current results. Noting that $\Delta V_{fg} = \Delta Q/C_T$, the achievable change in drain current due to programming relative to the initial drain current is given by

$$\frac{\Delta I}{I} \approx \frac{-\kappa \Delta Q}{U_T C_T} \tag{3}$$

where C_T is the total capacitance at the floating gate and ΔQ is the programmed charge. It is clear from (3) that the achievable precision is directly proportional to the charge that can be reliably transferred onto the floating gate and inversely proportional to the total floating-gate capacitance.

Similarly, for a floating-gate nFET operating in the strong inversion regime, the change in the drain current relative to the initial drain current is given by

$$\frac{\Delta I}{I} = \frac{2\kappa \Delta V_{fg}}{\kappa V_{fg} - V_s - \kappa V_{To}} = \frac{2\kappa \Delta Q}{V_{od} C_T} \tag{4}$$

where $V_{od} = \kappa V_{fg} - V_s - \kappa V_{To}$, is the overdrive voltage. The above equation is derived by assuming that the change in gate voltage due to programming is much smaller than the overdrive voltage of the device. As can be observed from (4), the achievable precision is directly proportional to the charge that can be transferred onto the floating-gate and inversely proportional to the overdrive voltage of the device and the total floating-gate capacitance.

In order to verify the theory presented above, a test chip was fabricated in a 0.5 μm standard CMOS process. Noting that for a transistor operating in the strong inversion regime an alternate expression for the overdrive voltage is $V_{od} = \sqrt{2\kappa I/K}$, it can be inferred from (4) that the FOM is inversely proportional to the square root of the drain current. Also, from (3), it is clear that the FOM is independent of drain current in the weak inversion regime. Therefore, one would expect that the plot of FOM versus drain current would be constant in the weak inversion regime and would increase in the strong inversion regime. This was verified by injecting a constant charge onto a floating-gate transistor and by measuring the *I–V* characteristic both before and after injection. Calculating the difference in currents between the *I–V* sweeps and plotting against the initial set of currents results in the plot shown in Fig. 6(a). As can be observed, the plot is constant in the weak inversion regime and increases in the strong inversion regime, thereby verifying the theory.

Table I presents quantitative numbers for the FOM for both the weak inversion and strong inversion regions based on the theory developed above. The FOM has been calculated for different values of charge transfer and C_T for a κ of 0.7, U_T of 26 mV, and an overdrive voltage of 250 mV. Fig. 6(b) presents

TABLE I
SUMMARY OF THE ACHIEVABLE BITS OF ACCURACY (FOM)

	Weak Inversion			Strong Inversion		
$\Downarrow C_T\ \Delta Q \Rightarrow$	1 **electron**	10 **electrons**	100 **electrons**	1 **electron**	10 **electrons**	100 **electrons**
10fF	11.18	7.85	4.83	13.44	10.12	6.8
100fF	14.5	11.18	7.85	16.76	13.44	10.12
1pF	17.82	14.5	11.18	20.09	16.76	13.44

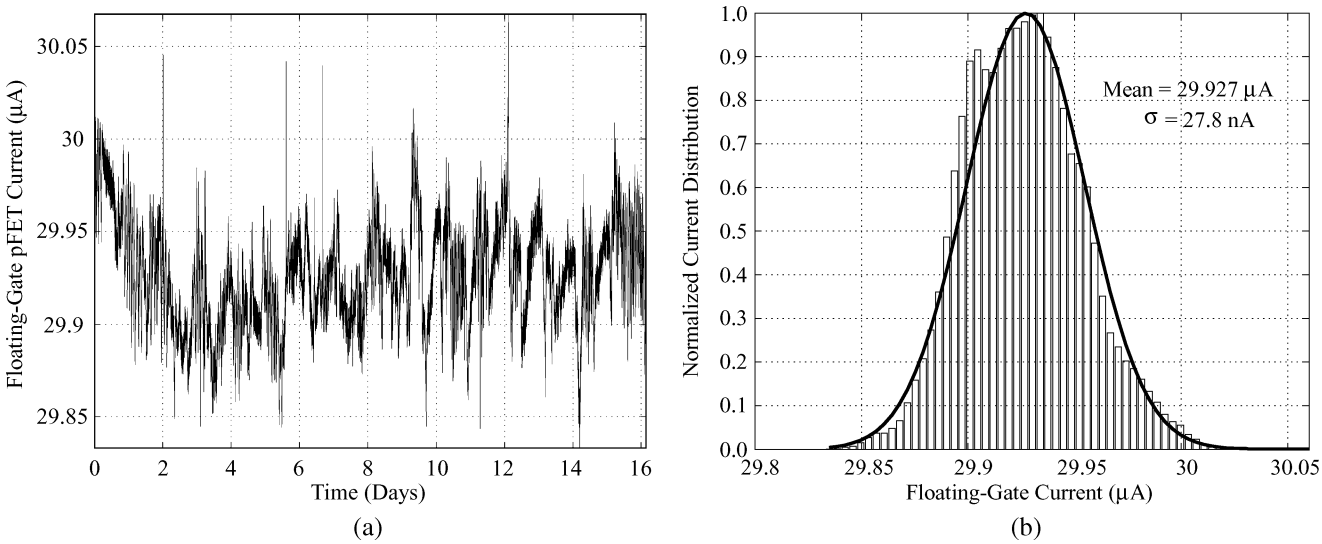

Fig. 7. Drain current of a floating-gate pFET. (a) The drain current of a floating-gate pFET measured over 16 days. The floating-gate transistor was programmed to an initial value of 30 μA. (b) The drain current distribution indicates a mean of 29.927 μA with a standard deviation of 27.8 nA. The Gaussian nature of distribution indicates the presence of thermal noise on the measured data.

experimental data from programming an array of floating-gate transistors to result in a sinusoid with a DC offset of 1 μA and an amplitude of 20 nA. Also shown is the percentage error in the programmed value of the sine wave to the ideal value. The error is within $\pm 0.05\%$, indicating an FOM of approximately 11 bits. The total floating-gate capacitance for these transistors is approximately 100 fF. With the devices operating in strong inversion, it can be inferred from Table I that a little over 100 electrons worth of charge is responsible for the measured precision. Using the above developed theory and depending on the region of operation of the floating-gate transistor, one can design a floating-gate transistor (W/L and C_T) such that a target accuracy specification is met.

IV. FLOATING-GATE CHARGE RETENTION

Floating-gate transistors inherently have good charge retention capabilities on account of the gate being surrounded by a high-quality insulator. Initial investigations of floating-gate retention were carried out by observing the drain current of a floating-gate device for long periods of time. Fig. 7(a) shows the drain current of a floating-gate pFET measured over a period of 380 hours. The drain current was programmed from a current of < 1 μA to an initial value of 30 μA and displayed a mean value of 29.93 μA with a standard deviation of 28 nA [see Fig. 7(b)]. The current exhibits a short-term drift in the beginning, beyond which no significant drift can be observed. This short-term drift is on account of the interface trap sites settling to a new equilibrium after programming [5]. Similar results have been observed in a 1.5 μm CMOS process [17]. Although this is a good indicator of the charge retention capabilities of floating-gate transistors, accurate estimates of the long-term charge retention can be made through accelerated lifetime tests.

Long-term charge loss in floating-gate transistors occurs due to the phenomenon of thermionic emission [4], [5], [18], [19]. The amount of charge lost is a function of both temperature and time and is given by

$$\frac{Q(t)}{Q(0)} = \exp\left[-t\upsilon \cdot \exp\left(\frac{-\phi_B}{kT}\right)\right] \tag{5}$$

where $Q(0)$ is the initial charge on the floating-gate, $Q(t)$ is the floating-gate charge at time t, υ is the relaxation frequency of electrons in polysilicon, ϕ_B is the Si-SiO_2 barrier potential in electron-volts, k is Boltzmann's constant, and T is the temperature. As expected from (5), charge loss in floating-gates is a slow process that is accelerated at high temperatures.

Floating-gate charge loss is measured indirectly by measuring the change in the transistor's threshold voltage. Programming floating-gate transistors by adding/removing charge modifies the threshold voltage of the device V_{To} as given by

$$V_{To} = V'_{To} + \frac{Q}{C_T} \tag{6}$$

where Q is the floating-gate charge, V'_{To} is the threshold voltage of the transistor with zero floating-gate charge or that of a non-floating-gate device, and C_T is the total capacitance at the gate node. Using the above approximation for the threshold voltage

TABLE II
SUMMARY OF FLOATING-GATE PARAMETER CHANGE IN 10 YEARS

	10% Programmed Change			50% Programmed Change		
Temperature	$\Delta Q/Q$	ΔV_{fg}	$\Delta I/I$	$\Delta Q/Q$	ΔV_{fg}	$\Delta I/I$
$25°C$	$10^{-3}\%$	$36.7nV$	$2\times10^{-4}\%$	$10^{-3}\%$	$156nV$	$9\times10^{-4}\%$
$90°C$	0.62%	$16.4\mu V$	0.06%	0.62%	$65\mu V$	0.47%
$140°C$	18.2%	$0.45mV$	1.8%	18.2%	$1.92mV$	10.7%

of a floating-gate device, the charge loss in a floating gate can be rewritten as

$$\frac{Q(t)}{Q(0)} = \frac{V_{To}(t) - V'_{To}}{V_{To}(0) - V'_{To}} \tag{7}$$

where $V_{To}(t)$ indicates the threshold voltage of the device after time t and $V_{To}(0)$ represents the initial programmed threshold voltage.

Estimating the amount of charge loss in floating-gate transistors requires the estimation of the parameters υ and ϕ_B as these parameters exhibit a wide spread in their values and therefore need to be extracted for each process. For the 0.5 μm process used in the design, floating-gate pFETs were programmed to a threshold voltage of -0.5 V and stored at high temperatures for a predefined time period. The change in threshold voltage is measured and using (7) the charge loss is estimated. Using (5), (7), and the measured data points, υ and ϕ_B can be extracted using

$$\phi_B = \frac{kT_1T_2}{T_1 - T_2} ln\left[\frac{t_2}{t_1}\cdot\frac{ln(x_1)}{ln(x_2)}\right] \quad \upsilon = \frac{-ln(x_1)}{t_1\cdot\exp\left(\dfrac{\phi_B}{kT_1}\right)} \tag{8}$$

where x denotes the ratio of the floating-gate charge at time t to the initial floating-gate charge and the subscripts denote two different data points. Using the above procedure, the values for the barrier potential and the relaxation frequency were extracted to be 0.9 eV and 60 s^{-1} for the 0.5 μm CMOS process used in the experiments.

Fig. 8 shows the measured floating-gate charge loss along with a theoretical extrapolated fit using the estimated model parameters. The measured data agrees well with the theoretical prediction and the trends observed in Fig. 8 have been observed across many floating-gate devices.

Now, consider two identical floating-gate transistors that have been programmed to a difference in current of ΔI, resulting in a differential floating-gate pair. Assuming weak inversion operation and noting that the analysis is similar to that in Section III, the difference in charge between the two floating gates is given by

$$\Delta Q = C_T\frac{U_T}{\kappa} ln\left(1 + \frac{\Delta I}{I}\right) = C_T\Delta V_{fg} \tag{9}$$

where all the variables have their usual meaning. Now, using (5) and the extracted values of ϕ_B and υ, the difference in charge at time t, namely $\Delta Q(t)$, can be estimated. From this, the difference in floating-gate voltage can be calculated, based on which and using (9), the value of the programmed difference current

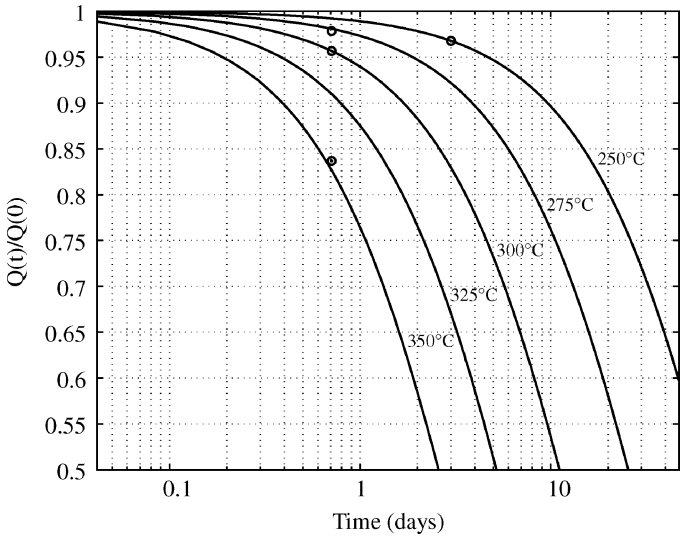

Fig. 8. Charge loss in floating-gate transistors. Charge loss measured at different temperatures and time periods as estimated from threshold voltage changes is plotted using o's. Parameters for a thermionic emission model were extracted using the measured data and the model is then used to calculate charge loss at different temperatures and time periods. This extrapolated theoretical fit is plotted using solid lines.

at time t ($\Delta I(t)$) can be estimated. Table II summarizes the data retention numbers for two different cases of programmed difference currents, namely, a 10% change and a 50% change for a time period of 10 years for different temperatures. A total floating-gate capacitance of 100 fF and a κ of 0.7 has been assumed for these calculations. As can be observed, floating-gate transistors display excellent charge retention capabilities.

The nonvolatile charge retention. when combined with programmability, makes floating-gate transistors well suited for use in precision analog circuits. Also, the analysis developed in Section III can be used in designing the aspect ratio and C_T of the floating-gate transistors such that a required programming precision is met. The application of a floating-gate pair with a difference current programmed between them to the design of an operational amplifier is considered next.

V. AMPLIFIER ARCHITECTURE

A single-stage folded cascode amplifier, shown in Fig. 9, demonstrates a practical implementation of the proposed approach shown in Fig. 1. The currents through the floating-gate transistor pair $M3$ and $M4$ are programmed such that they cancel the offset arising from mismatches in the input differential pair $(M1, M2)$ and the cascoded current mirrors $(M5-M8)$. During normal operation, the multiplexers $S1$ and $S2$ are set such that the floating-gate transistors are a part of the operational amplifier. During programming, the floating-gate transistors are isolated from the amplifier in order to program a

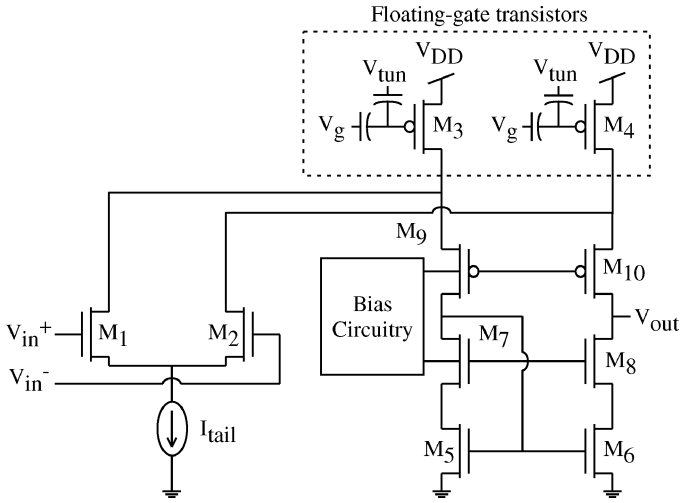

Fig. 9. Operational amplifier circuit schematic. A single-stage folded cascode amplifier that uses floating-gate transistors as trimming elements is shown. During normal operation switches S_1 and S_2 are set such that floating-gate transistors M_3 and $M4$ are a part of the operational amplifier. Offset voltage cancellation is achieved by programming a current difference between $M3$ and $M4$. Using floating-gate transistors both as a part of the amplifier and as trimming elements makes the architecture compact and easy to design.

difference current $\Delta I(I3-I4)$ such that the offset voltage is nullified.

A key advantage of this architecture is that the programming transistors are an integral part of the amplifier thereby simplifying the design process. Initially, all transistors including $M3$ and $M4$ are made non-floating-gate transistors and are designed to meet the amplifier's specifications. Next, these transistors $M3$ and $M4$ are made floating-gate transistors and, based on the offset requirement of the amplifier, an estimate can be made of the programming precision required. In other words, an approximate value of the difference current ΔI that needs to be programmed can be estimated, from which the FOM is calculated. Next, depending on the region of operation of the transistors $M3$ and $M4$, appropriate design equations developed in Section III can be used to estimate the total floating-gate capacitance needed. With the aspect ratio of the transistors set during the amplifier's design stage, the input capacitance and the tunneling capacitance can be sized to either meet or exceed the C_T requirement. Appropriate switches are then added to isolate the floating-gate transistors during programming. For this design, the floating-gate current sources were set to be 10 μA nominally and the total floating-gate capacitance was designed to be around 200 fF. From Table I it can be seen that a programming precision greater than 10 bits can be achieved for a charge transfer of around 100 electrons in strong inversion operation, which is sufficient for the design.

A. Input-Referred Offset Voltage

The amplifier exhibits zero offset voltage when all currents are balanced at its output. Assume that the amplifier has an uncompensated offset voltage given by V'_{off}. Let a current difference of ΔI_{fg} be programmed onto the pFET floating-gate transistors such that this difference current creates a voltage at the output equal to $\Delta I_{fg} r_o$, where r_o represents the effective output impedance at the output of the amplifier. The input-referred offset voltage of the amplifier therefore becomes

$$V_{\text{off}} = V'_{\text{off}} + \frac{\Delta I_{fg}}{g_{m1}} \tag{10}$$

where g_{m1} is the transconductance of the input differential pair. Therefore, based on (10), one would expect the input-referred offset voltage of the amplifier to exhibit a linear dependence with the programmed floating-gate difference current. Note that the above expression has been derived without assuming any specific region of device operation.

Now, performing a large-signal analysis by taking into account the mismatch between the various transistor pairs and referring their contributions to the input, it can be shown that for weak inversion operation, the input-referred offset voltage is given by

$$V_{\text{off}} = \Delta V_{th1} + \frac{I_3}{I_1}\Delta V_{fg} - \frac{I_9}{I_1}\Delta V_{th3} - \left(\frac{I_3}{I_1}\right)^2 \frac{\kappa}{U_T}\Delta V_{fg}^2 \tag{11}$$

and, similarly, for strong inversion operation, the offset voltage becomes

$$V_{\text{off}} = \Delta V_{th1} + \sqrt{\frac{I_3 K_3}{I_1 K_1}}\Delta V_{fg} - \sqrt{\frac{I_9 K_9}{I_1 K_1}}\Delta V_{th3}. \tag{12}$$

In the above equations, the threshold voltage mismatch between the input differential pair is accounted for with the threshold voltage of $M2$ being different from that of $M1$ by ΔV_{th1}. In the case of the transistor pair $M9/M10$, mismatch is represented by the threshold voltage of $M10$ being different from that of $M9$ by ΔV_{th3}. And, ΔV_{fg} represents the voltage difference between the floating-gates of $M4$ and $M3$ on account of the programmed charge difference between them. Note that both (11) and (12) simplify to (10) when the necessary small-signal approximations are made.

B. Temperature Sensitivity

Looking at (11) and (12), it is clear that the temperature sensitivity of the offset voltage can be estimated based on the sensitivities of the threshold voltage mismatch, ratios of transistor currents and K's. Note that ΔV_{fg} is temperature independent as for a typical operating temperature range, the charge loss on the floating-gate is negligible and therefore assumed constant, and to a first order, the total floating-gate capacitance is independent of temperature as well.

The temperature dependence of the threshold voltage is given by [16]

$$V_{th}(T) = V_{th}(T_o) + \alpha(T - T_o) \tag{13}$$

where T is the temperature in Kelvin, $V_{th}(T_o)$ represents the threshold voltage at a temperature T_o, and α represents the linear temperature coefficient of the threshold voltage. Now, the temperature dependence of the threshold mismatch between two devices can be written as

$$\Delta V_{th} = \Delta V_{th}(T_o) + \Delta\alpha(T - T_o) \tag{14}$$

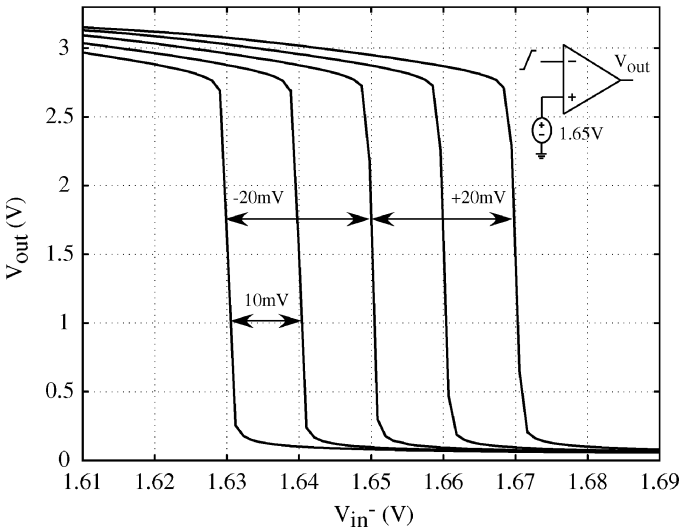

Fig. 10. Open loop DC transfer characteristics. The input offset voltage of the amplifier was programmed to five different values in steps of 10 mV. The non-inverting terminal of the amplifier was set at 1.65 V and the inverting terminal was swept from 0–3.3 V. The DC transfer curves show the switching points ranging from -20 to $+20$ mV with a 10 mV spacing as programmed.

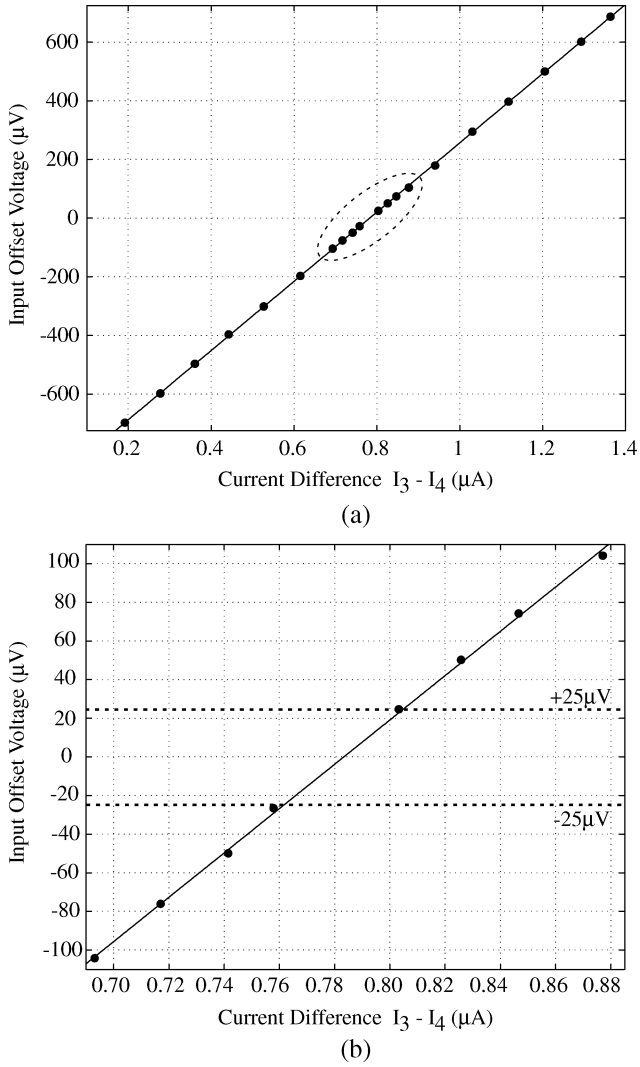

Fig. 11. Input offset voltage. (a) The input offset voltage of the amplifier was measured by programming different current differences between the floating-gate trimming transistors. The input offset voltage changes linearly with the difference current as expected from theory. (b) A zoom into the region of very low offset voltages. It is clear from the figure that offset voltages in the tens of microvolts are achievable, with the lowest being 25 μV.

where $\Delta V_{th}(T_o)$ represents the threshold mismatch at temperature T_o and $\Delta\alpha$ is the difference in their temperature coefficient.

The temperature dependence of κ has been analyzed by incorporating the temperature dependence of the terms describing κ and simulating using MATLAB. Assuming an n-channel transistor with a threshold voltage of 0.7 V with a temperature coefficient of -2 mV/°C, a substrate doping of 1×10^{17} cm^{-3}, a γ of 0.5 and a gate-bulk voltage of 1 V results in a κ of 0.8049 at room temperature (300 K). The variation of κ with temperature over a range of -40 °C to 140 °C was found to be ≈ 27 ppm/°C. Therefore, it was decided to assume κ to be constant with temperature to simplify the temperature analysis of the amplifier offset voltage.

Next, consider the term $\sqrt{IK}$ that appears in the expression for the input offset voltage in the strong inversion region of operation as given in (12). This term can be rewritten as

$$\sqrt{IK} = \mu_n C_{ox}\frac{W}{L}\left(\kappa(V_g - V_{To}) - V_s\right) = g_m \tag{15}$$

where all the terms are as defined earlier. Assuming fixed terminal voltages, the only terms that have a temperature dependence in the above equation are the threshold voltage and mobility. With regards to mobility, for two transistors of the same type, both the value and the temperature dependence can be assumed constant, thereby making the ratio temperature independent. For two transistors that are of dissimilar types, the electron and hole mobilities have a slightly different temperature dependence [20] and therefore result in a slight temperature dependent mobility ratio. However, for ease of analysis, the electron and hole mobilities will be treated as similar. With the above observations, and (14), the third term in (12) can be written as

$$\sqrt{\frac{I_9K_9(T)}{I_1K_1(T)}} = \left(\frac{K_9(T_o)\kappa V_{od9}(T_o)}{K_1(T_o)\kappa V_{od1}(T_o)}\right)\left(\frac{1-\dfrac{\kappa\alpha_9\Delta T}{V_{od9}(T_o)}}{1-\dfrac{\kappa\alpha_1\Delta T}{V_{od1}(T_o)}}\right) \tag{16}$$

where $\Delta T = T - T_o$. A similar expression can be arrived at for the second term in (12). Denoting $\kappa\alpha_1/V_{od1}(T_o)$ as a, $\kappa\alpha_9/V_{od9}(T_o)$ as b, $\kappa\alpha_3/V_{od3}(T_o)$ as c, and using (14) in (12) results in

$$V_{\text{off}}(T) = \Delta V_{th1} + \Delta\alpha_1\Delta T + \frac{gm_3(T_o)}{gm_1(T_o)}\frac{(1-b\Delta T)}{(1-a\Delta T)}\Delta V_g - \frac{gm_9(T_o)}{gm_1(T_o)}\frac{(1+c\Delta T)}{(1-a\Delta T)}(\Delta V_{th3} + \Delta\alpha_3\Delta T). \tag{17}$$

A similar analysis can be performed for weak inversion operation as well. As can be observed from (17), the offset voltage varies with the temperature and the variation can be approximated to be quadratic in nature. Also, it is clear that the offset voltage depends on threshold voltage mismatch multiplied by a ratio of quantities (transconductance). Since the threshold voltage mismatch by itself has a weak temperature

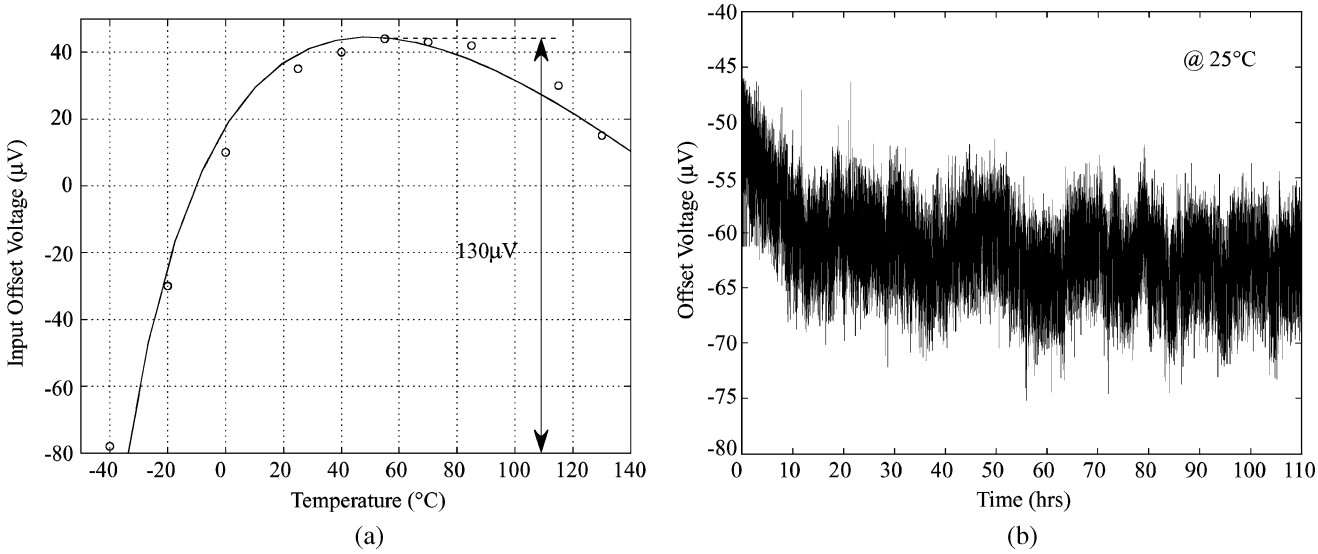

Fig. 12. Input offset voltage drift. (a) The input offset voltage of the amplifier was measured across a temperature range of −40 °C to 130 °C. The offset voltage displayed a maximum change of 130 μV across the entire temperature range. The o's represent the measured data points while the solid line represents the theoretical fit based on (17). (b) The input offset voltage of the amplifier was measured for a period of 110 hours at 25 °C. The offset was programmed from 1 mV to a reduced value of around −50 μV. It can be observed that after an initial short-term drift, the offset voltage drift is negligible.

dependence, designing the ratio of transconductances to be fairly temperature independent can result in an overall offset voltage that is temperature independent. This can be achieved by either biasing the transistors to their zero-temperature coefficient transconductances [21], [22] or by designing such that their overdrive voltages are close to each other, making the terms a, b, and c equal such that temperature sensitivity is minimized.

VI. Amplifier Experimental Results

Applying (10) and programming the drain currents of transistors $M3$ and $M4$, the amplifier was programmed to five different offset voltages. The offsets were programmed in steps of 10 mV ranging from −20 to +20 mV. Fig. 10 shows the DC transfer characteristics of the amplifier configured as a comparator with the non-inverting terminal held at 1.65 V. As can be observed, the comparator trip points are evenly spaced 10 mV apart as programmed. This clearly demonstrates the feasibility of the approach and the range of programming that is possible.

Accurate measurements of the offset voltage are made by using the amplifier under test along with a second amplifier configured as a nulling amplifier forming a servo loop [23]. Fig. 11(a) and (b) shows the measured input-referred offset voltage of the amplifier plotted against the various programmed floating-gate transistor difference currents. The measured data shows a linear dependence of the offset voltage with the programmed difference currents as expected from (10). As can be observed in Fig. 11(b), which zooms into the region encircled in Fig. 11(a), the offset voltage of the prototype amplifier can be programmed to 25 μV. Also, it can be seen that the amplifier can be programmed to display different offset voltages with both positive and negative polarities. This clearly demonstrates the programmable nature of the approach, a feature that could be exploited when designing, for instance, comparators. Experimentally, it is possible to program current increments as low as 0.1 nA. Theoretically, this indicates that offset voltages in the hundreds of nanovolts range are possible to achieve. At present, however, the primary limitation has been the internal noise of the amplifier itself.

Fig. 12(a) shows the sensitivity of the input offset voltage with temperature. The offset voltage was measured for temperatures ranging from −40 °C to 130 °C after programming at 25 °C. A maximum change of 130 μV was observed over the full temperature range of 170 °C. Since the transistors in the amplifier were biased in a region close to strong inversion, the temperature dependence was modelled according to (17). Shown in the figure is a theoretical fit of the data using (17). Since the exact values of the threshold voltage mismatch of the various transistor pairs are unknown, the fit was performed using a reasonable set of parameter values. It should be noted that the exact shape of the temperature characteristic depends on the transistor operating regions, biasing conditions, and the mismatch between threshold voltages.

In order to experimentally observe the offset drift with time, the amplifier was programmed to an initial offset voltage of around −50 μV from an initial offset voltage of 1 mV and measured continuously over a period of 110 hours. Fig. 12(b) shows the measurement of the offset voltage with time. As can be observed, the offset voltage exhibits an initial short term drift of about −10 μV on account of the interface trap sites settling to a new equilibrium. Beyond the initial short-term drift, the offset voltage drift is negligible, as expected from earlier measurements on floating-gate charge retention.

Table III summarizes the performance of the amplifier and the chip micrograph is shown in Fig. 13. The total area of the amplifier excluding the buffer is 115 μm × 45 μm and the additional area occupied by the input capacitors and the switches on account of using floating-gate transistors is 45 μm × 45 μm. As can be observed, using floating-gate transistors as a part of the amplifier and also as a programming element leads to a compact

TABLE III
OPERATIONAL AMPLIFIER SUMMARY OF PERFORMANCE

Parameter	Value
Supply Voltage	$3.3V$
Technology	$0.5\mu m$ CMOS
Input Common Mode Range	$1.2V - 3.1V$
Output Voltage Swing	$0.2V - 3.1V$
Input Offset Voltage	$\pm 25\mu V$
Offset Voltage Drift with Temperature	$130\mu V/170°C$
Offset Voltage Drift @ $55°C$ for 10 yrs	$< 0.5\mu V$
Open Loop Gain	$63dB$
Unity Gain Bandwidth @ $C_L = 20pF$	$10MHz$
Phase Margin	$60°$
Common Mode Rejection Ratio	$73dB$ (Simulation)
Power Supply Rejection Ratio	$77dB$ (Simulation)
Input Referred Noise (rms)	$8.9\mu V$ (Simulation)
Slew Rate	$5V/\mu s$
Settling Time (10 Bit) for $100mV$ Step	$105ns$
Power Dissipation (Excl. Buffer)	$66\mu W$
Power Dissipation (Incl. Buffer)	$8.25mW$
Area (Excl. Buffer)	$115\mu m \times 45\mu m$

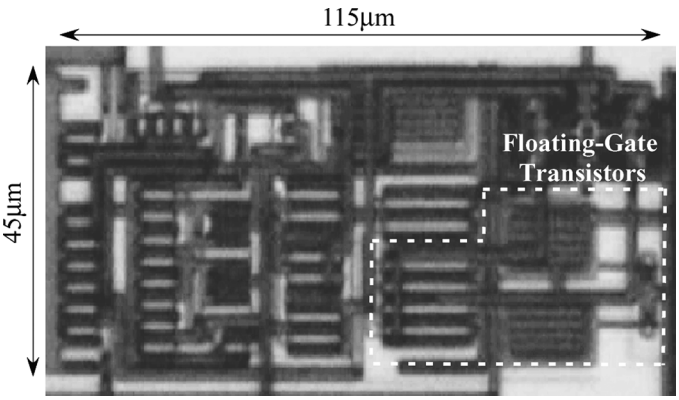

Fig. 13. Amplifier die micrograph. The chip micrograph of the prototype operational amplifier excluding the output buffer is shown to occupy an area of 115 μm × 45 μm. The additional area on account of using floating-gate transistors is 45 μm × 45 μm.

architecture. Also, the proposed cancellation scheme is independent of other amplifier parameters.

Automatic programming of the floating-gate transistor makes the approach attractive from a commercial standpoint. Unlike wafer trimming, which is susceptible to offset drifts because of packaging stress, the proposed scheme involves offset cancellation at the package level. Extra pins (V_{tun}, V_g, V_{drain}) and digital pins for the drain selection circuitry are needed for programming multiple floating-gate transistors. The programming infrastructure allows the gate, drain, and tunnel voltages to be shared among different floating-gate transistors. This keeps the number of extra pins required constant even when using multiple floating-gate transistors, a scenario that is typical while using multiple amplifiers on the same chip. The pin count can be reduced if the gate voltage is supplied by the biasing structure during programming as well and by using a serial digital interface for the digital pins.

The reduction of the offset voltage involves three main steps: 1) measurement of the offset voltage; 2) programming of the floating-gate transistor; and 3) switching between the measurement and the programming mode. Typical convergence to minimum offset voltage takes about 10 programming pulses of a 100 μs each. A more detailed discussion can be found in Section II. The programming time of minimum offset voltage is determined by the measurement speed, as this is the slowest of the steps by far. Typical programming times for the minimum offset voltage is on the order of 10–20 s.

VII. COMPARISON TO PREVIOUS WORK

The use of floating-gate transistors to correct for mismatch in analog circuitry has been investigated by other authors as well [4], [5]. The approach in [4] results in an unidirectional offset cancellation. This requires an intentional offset creation of the correct polarity during the design phase of the amplifier for proper operation. This intentional offset creation has been cited as the reason for the degradation of the offset voltage temperature sensitivity [4]. The work in [5] introduces a trimming circuitry based on floating-gate transistors to produce a difference current which is then used as a building block to compensate for mismatch induced errors. The proposed approach in this paper is conceptually similar to that in [5] in that it uses a differential current to trim offsets. However, the difference current is created using just two floating-gate transistors which then form an integral part of the amplifier of interest. This results in an advantage in terms of both area and design overhead. Also, the proposed approach uses hot-electron injection to program floating-gate transistors while both [4] and [5] use Fowler–Nordheim tunneling as the primary programming mechanism. The advantages of an injection based programming scheme over a tunneling based programming has been highlighted earlier in Section II.

Correcting analog circuit mismatch using resistor trimming is an alternate technique. Resistor trimming is usually performed using laser annealing, laser trims, poly fuses and zener zapping. Both laser annealing and laser trims are expensive and do not provide the flexibility of in-package trims. Trimming using poly fuses and zener zapping is discrete in nature and therefore accuracy is limited to the smallest resistor step used. Also, using a number of zener diodes and poly fuses involves an area penalty. All of the above resistor trimming techniques are one-time programmable. The approach described in this paper is cost effective, field programmable and is a package level correction scheme.

The proposed approach involves lesser design overhead when compared to the technique of using current-mode digital-to-analog converters controlled using an EEPROM and serial interface [2], [22] to reduce amplifier offsets. Also, the proposed approach can provide a continuous range of offset voltages rather than discrete values offered by the DAC-based scheme. This makes the approach well suited for other applications as well, such as, programming a chain of comparators to different trip points for use in, say, Flash analog-to-digital converters.

Auto-zeroing is primarily useful for sampled data systems and is limited by issues such as charge injection, clock feedthrough, and wideband noise folding into the baseband on account of undersampling. For a continuous-time operation, chopper stabilization or continuous-time auto-zeroing such as a ping-pong amplifier [24] are the typical alternatives. The chopper amplifier is, however, limited in use to low-bandwidth applications [1]. The ping-pong approach involves the

TABLE IV
COMPARISON OF OFFSET CANCELLATION SCHEMES

	FGate	**Autozero**	**Chopper**	**Ping-Pong**	**R Trimming**	**DAC**
Mode	Continuous	Sampled	Continuous	Continuous	Continuous	Continuous
Offset (V_{os})	Low	Moderate	Low	Moderate	Low	Low
Bandwidth	High	High	Low	High	High	High
Complexity	Low	Moderate	High	Moderate	Moderate	Moderate
$1/f$ Noise	No effect	Reduced	Reduced	Reduced	No effect	No effect
Extra Power	Low	Moderate	Moderate	Moderate	Low	Moderate
Extra Area	Low	Moderate	Moderate	Moderate	Moderate	High
V_{os} Removal	Long-Term	Periodic	Continuous	Periodic	Long-Term	Long-Term
Field Programmability	Yes	No	No	No	No	Yes

use of multiple amplifiers and multi-phase clocks that add additional overhead in terms of area and power. The proposed floating-gate approach involves none of the above tradeoffs and the offset cancellation by itself dissipates no additional power. The approach places minimal overhead on the amplifier design with nonvolatile storage of offset reduction information. The primary limitation, however, is the lack of flicker noise reduction. Finally, Table IV summarizes qualitatively the design tradeoffs of the proposed approach to the various offset cancellation schemes on the different design parameters of interest.

VIII. CONCLUSION

An amplifier topology has been presented that uses floating-gate programmable elements as an integral part of the amplifier. The approach places minimal overhead on the amplifier design with nonvolatile storage of offset reduction information. A prototype amplifier has been fabricated in a 0.5 μm standard digital CMOS process and trimmed to an offset voltage of 25 μV. The offset voltage exhibits a temperature sensitivity of 130 μV over a temperature range of 170 °C. Floating-gate transistors surrounded completely by SiO_2, a high-quality insulator, exhibit excellent charge retention capabilities. A long-term offset voltage drift of less than 0.5 μV when stored at a temperature of 55 °C for 10 years has been calculated based on a thermionic emission model for floating-gate charge loss. Direct tunneling through the gate oxide (gate leakage) is a limitation for charge retention in floating-gate transistors for oxide thicknesses less than 5 nm which are typical for finer line processes (< 0.25 μm). However, the proposed approach is still scalable with process technologies. Floating-gate transistors have not been used in the signal path, therefore in smaller dimension processes, floating-gate transistors can be implemented using the available thick oxide transistors with no impact on the speed of operation of the amplifier. Using thicker oxides preserve the charge retention capability of floating-gate devices thereby providing low long-term drifts in the amplifier offset voltage. Finally, programmability coupled with a negligible long-term drift and scalability makes this approach attractive for offset reduction in operational amplifiers.

ACKNOWLEDGMENT

The authors would like to thank K. Sundaresan, Dr. F. Ayazi, Dr. A. Doolittle, and Dr. P. Patel of Georgia Institute of Technology, Atlanta, GA, for help with the temperature measurements. The authors would also like to thank Dr. Y. Tsividis for his encouragement and many valuable discussions.

REFERENCES

[1] C. C. Enz and G. C. Temes, "Circuit techniques for reducing the effects of op-amp imperfections: Autozeroing, correlated double sampling, and chopper stabilization," *Proc. IEEE*, vol. 84, no. 11, pp. 1584–1614, Nov. 1996.

[2] W. J. Kim, S. Sompur, and Y. B. Kim, "A novel digial controlled technique for operational amplifier compensation," in *Proc. MidWest Symp. Circuits and Systems*, Aug. 2001, pp. 211–214.

[3] G. Erdi, "Never-mentioned op amp issues," in *Proc. Bipolar Circuits and Technology Meeting*, Minneapolis, MN, Sep. 1990, pp. 219–222.

[4] L. R. Carley, "Trimming analog circuits using floating-gate analog MOS memory," *IEEE J. Solid-State Circuits*, vol. 24, no. 6, pp. 1569–1575, Dec. 1989.

[5] E. Sackinger and W. Guggenbuhl, "An analog trimming circuit based on a floating-gate device," *IEEE J. Solid-State Circuits*, vol. 23, no. 6, pp. 1437–1440, Dec. 1988.

[6] F. Adil, G. Serrano, and P. Hasler, "Offset removal using floating gate circuits for mixed-signal systems," in *Proc. Southwest Symp. Mixed-Signal Design*, Feb. 2003, pp. 190–195.

[7] D. Kahng and S. M. Sze, "A floating-gate and its application to memory devices," *Bell Syst. Tech. J.*, vol. 46, no. 4, pp. 1288–1295, 1967.

[8] S. Lai, "Flash memories: Where we were and where we are going," in *IEDM Tech. Dig.*, San Francisco, CA, 1998, pp. 971–973.

[9] M. Lezlinger and E. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," *J. Appl. Phys.*, vol. 40, no. 1, pp. 278–283, Jan. 1969.

[10] S. Kinoshita, T. Morie, M. Nagata, and A. Iwata, "A PWM analog memory programming circuit for floating-gate MOSFETs with 75-μs programming time and 11-bit updating resolution," *IEEE J. Solid-State Circuits*, vol. 36, no. 5, pp. 1286–1290, May 2003.

[11] W. Gao and W. M. Snelgrove, "Floating gate charge-sharing: A novel circuit for analog trimming," in *Proc. IEEE ISCAS'94*, May 1994, pp. 315–318.

[12] C. Duffy and P. Hasler, "Modeling hot-electron injection in PFETs," *J. Comput. Electron.*, vol. 2, no. 2-4, pp. 317–322, Dec. 2003.

[13] G. Serrano, P. Smith, H. Lo, R. Chawla, T. Hall, C. Twigg, and P. Hasler, "Automatic rapid programming of large arrays of floating-gate elements," in *Proc. IEEE ISCAS 2004*, May 2004, pp. 373–376.

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

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

[16] Y. Tsividis, *Operation and Modeling of the MOS Transistor*. New York: McGraw-Hill, 1999.

[17] B. K. Ahuja, H. Vu, C. L. Aber, and W. Owen, "A 0.5 μA precision CMOS floating-gate analog reference," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2005, pp. 286–287.

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

[19] H. Nozama and S. Kokyama, "A thermionic electron emission model for charge retention in SAMOS structures," *Jpn. J. Appl. Phys.*, vol. 21, pp. L111–L112, Feb. 1992.