Characterization, Modeling, and Minimization of Transient Threshold Voltage Shifts in MOSFET's

Theodore L. Tewksbury, III, *Member, IEEE*, and Hae-Seung Lee, *Senior Member, IEEE*

Abstract— **MOSFET's subjected to large-signal gate–source voltage pulses on microsecond to millisecond time scales exhibit transient threshold voltage shifts which relax over considerably longer periods of time. This problem is important in high-accuracy analog circuits where it can cause errors at the 12 b level and above. In this paper, transient threshold voltage shifts are characterized with respect to their dependence on stress amplitude and duration, relaxation time, gate bias, substrate bias, drain voltage, temperature, and channel width and length. In contrast to previous studies, threshold voltage shifts are measured at time and voltage scales relevant to analog circuits, and are shown to occur even when the effects of Fowler–Nordheim tunneling, avalanche injection, hot carriers, trap generation, self-heating, mobile ions, and dipolar polarizations are absent. A new model is proposed in which channel charge carriers tunnel to and from near-interface oxide traps by one of three parallel pathways. Transitions may occur elastically, by direct tunneling between the silicon band edges and an oxide trap, or inelastically, by tunneling in conjunction with a thermal transition in the insulator or at the Si–SiO_2 interface. Simulations based on this model show excellent agreement with experimental results. The threshold voltage shifts are also shown to be correlated with $1/f$ noise, in corroboration of the tunneling model. Techniques for the minimization and modeling of errors in circuits are presented.**

I. Introduction

In MOS analog circuits such as switched-capacitor filters [1], [2], digital-to-analog (D/A), and analog-to-digital (A/D) converters [3], [4], the differential input stages of comparators and operational amplifiers are routinely subjected to momentary asymmetrical voltage stresses during the course of their normal operation. In comparators, the magnitude of this stress can be as large as half the full scale voltage of the A/D converter, while in operational amplifiers, it can be on the order of the supply voltage under slewing conditions. MOSFET's subjected to these large-signal stress conditions exhibit transient threshold voltage shifts with amplitudes up to 1 mV and relaxation times comparable to, or longer than, clock periods in typical analog circuits [5], [6], causing linearity errors, long-settling tails, and hysteresis effects. In a previous study [5], we measured transient threshold voltage shifts in MOS differential pairs, but the range of bias conditions was limited and a complete theoretical explanation has not yet been offered. This paper extends our previous work to include measurements of transient threshold voltage shifts over a wide range of conditions relevant to analog circuits, and presents a new model to explain the results.

Threshold voltage shifts in MOS transistors have been studied extensively, particularly in relation to device lifetime and reliability. However, most of these studies [7]–[9] have focused on long time scales and high voltages, far outside the range of interest for analog circuits. Under these conditions, shifts in device characteristics can be caused by a variety of mechanisms, including the drift of mobile ions [10], the polarization of molecular dipoles [8], [10], self-heating [11], and charge trapping by interface states and oxide traps [9], [12]. Under high vertical fields ($>$ 7 MV/cm), traps can be created in the insulator [13] and can be charged by Fowler–Nordheim tunneling [14] or avalanche injection [15] of carriers into the oxide bands. At high drain voltages, channel hot carriers can generate interface states [16] and can be injected into the oxide, causing threshold voltage instabilities and relaxation effects [17], [18]. At the lower voltages and shorter time scales characteristic of normal circuit operation, channel charge carriers can be exchanged with oxide traps by direct tunneling through the interface potential barrier [12], [19]. While this mechanism is known to be associated with the $1/f$ noise in MOS devices [20], its manifestation under large-signal stress conditions has not yet been adequately investigated. This paper reports on a new measurement technique which isolates direct tunneling from competing effects and enables this mechanism to be characterized under large-signal stress conditions.

Existing theoretical models of charge trapping by direct tunnel exchange with oxide traps [7], [9], [19]–[21] are inadequate to explain the observed threshold voltage transients due to several deficiencies. First, many of the assumptions on which these models are based are violated under large-signal stress. For example, band bending in the oxide is normally neglected and Boltzmann statistics are assumed, even though the Fermi level is driven well into the silicon bands. Second, despite the fact that it is an elastic process, tunneling is commonly assumed to occur between the silicon bands and oxide traps at arbitrary energies within the bandgap, without a plausible explanation for how the required change in energy takes place [22]. Third, we have observed deviations from the strict logarithmic time dependence derived by previous researchers [7], [23], and the voltage and temperature dependence of our measurements cannot be explained by existing models. New theory is presented here which overcomes these problems and

Manuscript received June 30, 1993; revised October 19, 1993. This work was supported in part by the Semiconductor Research Corporation (SRC) under Contract 92-SP-309 and by the National Science Foundation (NSF) under Contract M1P8858020.

T. L. Tewksbury, III is with Analog Devices Semiconductor, Wilmington, MA 01887.

H.-S. Lee is with the Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139.

IEEE Log Number 9214797.

enables transient threshold voltage shifts to be quantitatively explained and modeled.

Section II begins with a description of the measurement technique and presents the key experimental observations which suggest a tunneling mechanism. Section III then presents a physically based model which is consistent with the inherently elastic nature of tunneling and includes band bending in the oxide, Fermi–Dirac statistics, nonuniform trap densities in energy and space, and correlated mobility fluctuations. An approximate solution for the threshold voltage shift is derived, and the well-known logarithmic time dependence is shown to be a limiting case of a more general relation. In Section IV, transient threshold voltage shifts are characterized and simulated as a function of stress amplitude and duration, relaxation time, gate bias, substrate bias, drain voltage, temperature, and device size. Finally, Section V discusses scaling issues, differences between NMOS and PMOS transistors, the relationship between threshold voltage shifts and $1/f$ noise, and techniques for the minimization and modeling of errors in circuits.

II. Experimental Procedure

A. Measurement Circuit

Matched pairs of integrated circuit MOSFET's were packaged and tested using the circuit of Fig. 1(a). Before and after stress, with the CMOS switches in the positions shown (ϕ_1 high, ϕ_2 low), the source-coupled pair comprises the input stage of a noninverting feedback amplifier with gain $(R_1 + R_2)/R_2 \approx 100$ from the gate of the device under test ($M1$) to the output. A precision voltage supply in the sources sets the initial gate–source bias, $V_{GS} = -V_{\text{BIAS}}$. Offset voltage is nulled by adjustment of resistors R_3 and R_4 so that the gate of $M1$ is initially at ground. During stress (ϕ_1 low, ϕ_2 high), the gate of $M1$ is switched to a voltage source V_p and the opamp is disconnected from the feedback loop. This prevents stressing of the opamp input stage which could interfere with the measurement due to thermal or charge-trapping hysteresis [5]. Following stress for a period $1\ \mu\text{s} < t_s < 10$ ms, the switches are returned to their original positions. Feedback returns the drain current of $M1$ to its initial value by producing an *equivalent gate–source voltage shift* that precisely compensates for any change in threshold voltage or mobility resulting from stress. The amplified gate–source voltage transient is averaged and observed at the output V_{out} on a digitizing oscilloscope. The design ensures that the output remains within several millivolts of ground at all times during the measurement in order to prevent overdriving of the oscilloscope preamplifier, which could otherwise contribute long-settling components to the response. The output is Schottky-clamped in order to enforce this condition, even under transient conditions.

MOSFET's are biased in the linear region at low drain–source voltage ($|V_{DS}| < 500$ mV) so that channel hot carrier generation is negligible. Vertical electric fields are maintained less than 1.5 MV/cm to prevent trap generation and injection of carriers into the oxide bands. In order to eliminate self-heating, the device is biased such that no change in power

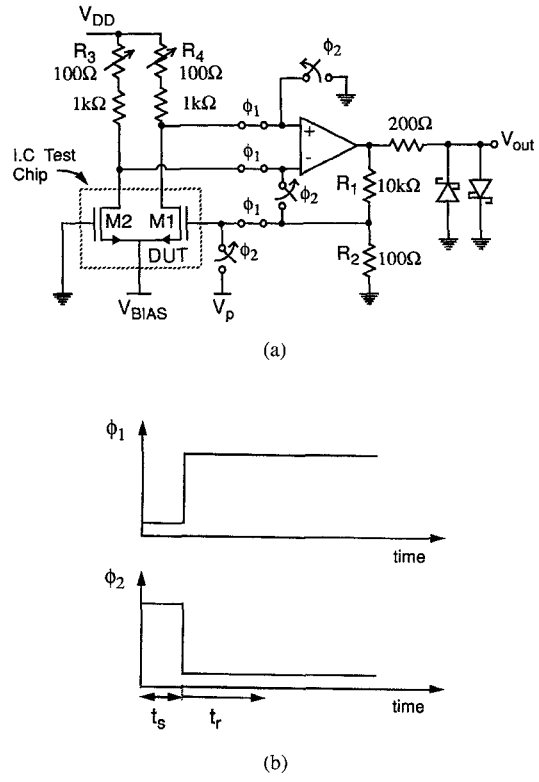


Fig. 1. Measurement circuit. (a) Schematic. (b) Clock signals for CMOS switches.

dissipation occurs during the measurement sequence:

$$I_{DS1}V_{DS1} = I_{DS2}V_{DS2} = I_{DS3}V_{DS3} = \text{constant} \qquad (1)$$

where the subscripts 1, 2, and 3 indicate the bias points of $M1$ before, during, and after stress, respectively.

B. Device Fabrication

Threshold voltage shifts were characterized on transistors from five different CMOS and BiCMOS processes with 250–750 Å thermally grown SiO_2 gate dielectrics, as summarized in Table I. Samples 1–3 are n-well BiCMOS processes employing a wet–dry–wet gate oxidation and a blanket boron threshold voltage adjustment implant. Sample 4 is a p-well CMOS process with no threshold implant and a dry gate oxidation followed by N_2 anneal. Process 5 is a twin-well CMOS technology with separate n and p threshold implants and a dry gate oxidation followed by a N_2 anneal.

C. Basic Observations

Fig. 2 shows typical threshold voltage relaxation transients measured on an NMOS transistor stressed for 10 ms at various voltages. At $V_p = +5$ V, the initial threshold voltage shift is about 1 mV and decays over a period exceeding 100 ms. While too small to be of interest in most digital applications, threshold voltage shifts of this amplitude are sufficient to cause errors at the 12 b level in an A/D converter with a 5 V reference. Table I summarizes measurements from each of the five processes. The single number used as a figure of merit is the amplitude of the threshold voltage shift sampled

TABLE I
STATISTICS OF THRESHOLD VOLTAGE SHIFTS IN FIVE DIFFERENT PROCESSES (t_s = 10 ms, t_r = 1 ms)

Process	t_{ox} (Å)	W/L	NMOS $\Delta V_T(\mu V)$						PMOS $\Delta V_T(\mu V)$					
			V_{GS} (V)	V_{DS} (mV)	$V_p = +5$ V		$V_p = -5$ V		V_{GS} (V)	V_{DS} (mV)	$V_p = +5$ V		$V_p = -5$ V	
					mean	σ	mean	σ			mean	σ	mean	σ
1	350	20/5	1.0	200	370.0	50.0	−135.0	22.4	−1.5	−500	27.2	26.1	−5.3	3.0
2	350	100/5	1.0	200	700.1	113.5	−368.6	189.1	−1.5	−500	37.3	28.1	−20.0	11.9
3	750	48/4	1.2	300	482.8	19.4	−135.0	44.6	−2.0	−500	37.2	10.0	−49.6	29.5
4	500	20/5	1.0	200	65.2	28.0	−33.8	21.8	−2.0	−500	11.9	2.3	−4.8	1.8
5	250	20/5	1.2	300	523.6	72.2	−432.8	236.0	−1.4	−500	134.6	93.4	−28.7	17.3

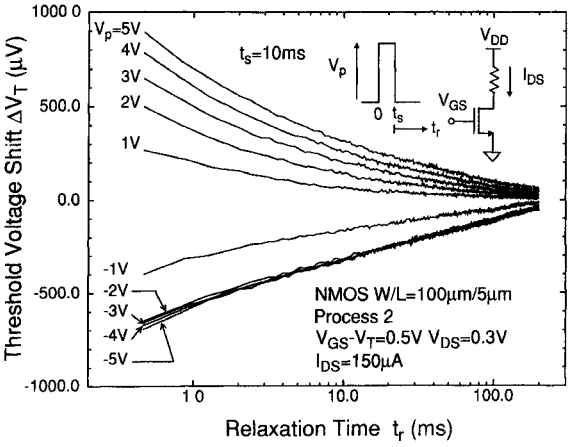


Fig. 2. Measured threshold voltage relaxation transients in an NMOS transistor stressed for 10 ms at various voltages.

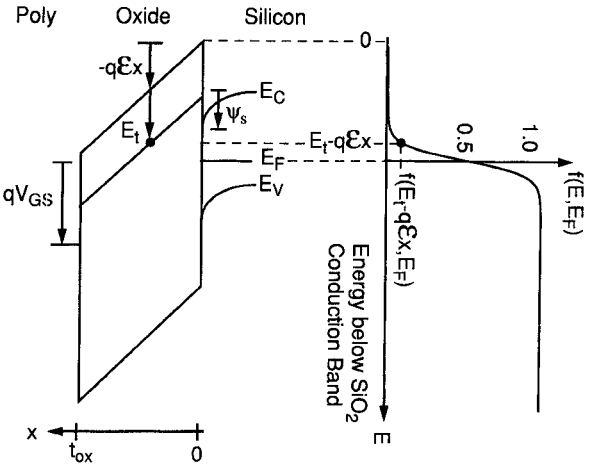


Fig. 3. Energy band diagram of an NMOS transistor showing definition of coordinates ($V_{BS} = 0$).

a fixed delay, $t_r = 1$ ms, after stress for $t_s = 10$ ms. The peak threshold voltage shift may be much larger, but cannot be measured due to the finite response time of the circuit. For each process, the mean threshold voltage shift is calculated from measurements on a randomly chosen sample of 10–20 devices from the same lot.

Several observations can be made from these data. 1) The threshold voltage shifts cannot be caused by hot carriers, trap generation, self-heating, avalanche, or Fowler–Nordheim injection since the measurement conditions have been specifically designed to preclude these effects. 2) The polarity of the threshold voltage shifts is always in the same direction as the stress voltage, $\Delta V_T/V_p > 0$, opposite to that which would be produced by mobile ions or dipolar polarizations [10]. 3) The large inter- and intraprocess variations of the threshold voltage shifts are consistent with a mechanism that depends on a random distribution of defects. However, the time constants of the decay are too long to be explained entirely by fast surface states at the Si–SiO_2 interface. These observations, together with the nearly logarithmic time dependence of the relaxation, suggest a mechanism in which channel charge carriers are exchanged with oxide traps by direct tunneling [7], [23].

III. TUNNELING MODEL

A. Kinetics

Fig. 3 shows the band diagram of an NMOS transistor with the definition of coordinates used in the following analysis. The energy level of an oxide trap is specified by its depth E_t below the SiO_2 conduction band. Consistent with this definition, and contrary to the usual convention, the energy of an electron is specified by its depth E below the SiO_2 conduction band at the interface ($x = 0$). The energy E of an oxide trap therefore shifts under an applied oxide electric field $\mathcal{E}$ by an amount $q\mathcal{E}x$:

$$E = E_t - q\mathcal{E}x. \tag{2}$$

Note that, with distance x increasing leftward into the oxide in Fig. 3, $\mathcal{E} < 0$ for $V_{GS} > 0$. Oxide traps near the interface eventually reach equilibrium with a fractional occupancy governed by the Fermi–Dirac distribution function

$$f(E, E_F) = \frac{1}{1 + \exp[(E_F - E)/kT]}$$

where E_F is the Fermi level at the surface.

Fig. 4 illustrates the proposed tunneling model for the case of an NMOS transistor subjected to positive stress. Before stress, the electric field in the oxide is $\mathcal{E}_o = -V_{ox}/t_{ox}$ and the bands are bent as shown in Fig. 4(a). Traps below the initial Fermi level ($E > E_{F0}$) are mostly filled, while those above ($E < E_{F0}$) are mostly empty. During stress, the Fermi level at the interface rises relative to the conduction band, and oxide trap energy levels are lowered by the field $\mathcal{E}$ in the insulator. Traps fill to a level determined by the new Fermi level E_F as shown in Fig. 4(b). When the stress is removed, trap energies return to their original levels and trapped carriers tunnel back out to the interface [Fig. 4(c)]. Assuming first-order kinetics [19], [24], the change in the density of occupied

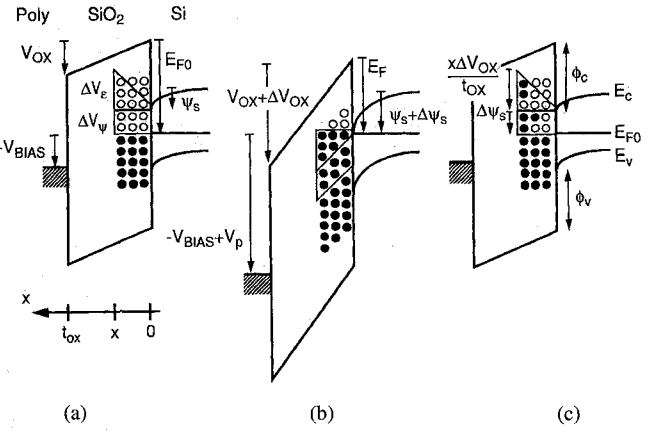


Fig. 4. Energy band diagrams showing windows in E–x space containing traps that contribute to a threshold voltage shift ($V_{BS} = 0$). (a) Prestress; traps empty. (b) During stress; traps capture electrons. (c) After stress; trapped electrons tunnel out.

oxide traps δn_{ot} ($\text{cm}^{-3} \cdot \text{eV}^{-1}$) at energy E_t and depth x from the interface is

$$\delta n_{ot}(E_t, x, t_s, t_r) = D_{ot}(E_t, x)\Delta f(E_t, x)(1 - e^{-t_s/\tau_i})e^{-t_r/\tau_j} \quad (3)$$

where

$$\Delta f(E_t, x) = f(E_t - q\mathcal{E}x, E_F) - f(E_t - q\mathcal{E}_o x, E_{F0}) \quad (4)$$

with $D_{ot}(E_t, x)$ ($\text{cm}^{-3} \cdot \text{eV}^{-1}$) the oxide trap density, t_s the stress time, and t_r the relaxation time following stress. The time constants during stress (τ_i) and during emission (τ_j) depend on the details of the trapping mechanism as described below.

B. Time Constants

Tunneling is an elastic process requiring the presence of occupied states on one side of a potential barrier and unoccupied states at the same energy on the other side. Due to band bending in the oxide, this requires that the energy of the tunneling electron E be related to the energy level E_t of the oxide trap into which it tunnels by (2). The time constant for the tunneling transition is given by [24], [25]

$$\tau_{\text{tun}}(E_t, x) = \tau_o(E_t, x)e^{\lambda(E_t - q\mathcal{E}x, x)} \quad (5)$$

with

$$\lambda(E, x) = \begin{cases} \dfrac{4}{3}\left(\dfrac{2m^*}{\hbar^2}\right)^{1/2} \dfrac{(E + q\mathcal{E}x)^{3/2} - E^{3/2}}{q\mathcal{E}} \\ \quad |\mathcal{E}| > 0 \,(\text{trapezoidal barrier}) \\ 2\left(\dfrac{2m^*E}{\hbar^2}\right)^{1/2} x = 2Kx \\ \quad \mathcal{E} = 0 \,(\text{rectangular barrier}) \end{cases} \quad (6)$$

where m^* is the effective mass of the tunneling carrier in the oxide. The prefactor τ_o is weakly dependent on energy and depth [26] and can be assumed constant, with a value that depends on whether the tunneling transition occurs from states in the conduction band, in the valence band, or within the bandgap.

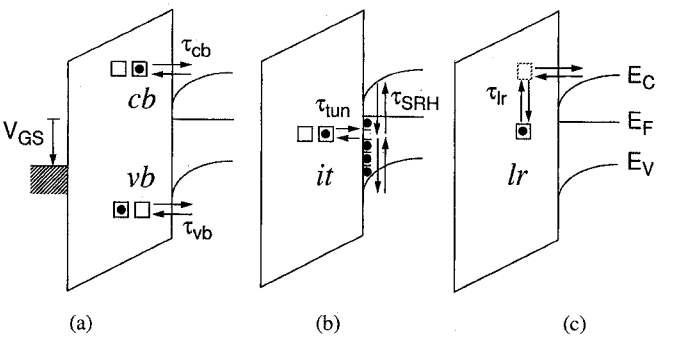


Fig. 5. Three pathways for exchange of channel charge carriers with oxide traps. (a) Elastic tunneling from the conduction band [cb] and valence band [vb]. (b) Two-step (interface trap assisted) process [it]. (c) Tunneling followed by lattice relaxation multiphonon emission [lr].

Electrons above the silicon conduction band can tunnel into oxide traps at the same energies as shown in Fig. 5(a). The time constant τ_{cb} for this process is given by (5) with $E \approx \phi_c = 3.1$ eV, $m^* = m_e^* \approx 0.42m_o$ [26], and $\tau_o \approx 10^{-10}$ s [20]. The exponential dependence of (6) on trap depth leads to a broad distribution of time constants which can readily explain the long observed relaxation times. Using the above values, a trap 20 Å deep in the oxide has a time constant on the order of 1 s, so that traps which change occupancy during the measurement are confined to a very thin layer near the interface. An expression analogous to (6) can be written for the time constant τ_{vb} of holes tunneling from the valence band. However, due to their larger effective mass [$m_h^* \approx (10$ –$20)m_e^*$] [27] and the higher potential barrier at the valence band ($\phi_v \approx \phi_c +$ 1 eV), the time constant for holes is much greater than for electrons.

Charge exchange can also occur between the silicon bands and oxide traps at energies within the bandgap, but this requires the assistance of an additional mechanism to supply or dissipate the required change in energy, a point overlooked in much of the literature. One mechanism by which this can occur is shown in Fig. 5(b). In this two-step process, channel charge carriers communicate thermally with interface traps (fast surface states) which, in turn, exchange charge with oxide traps at the same energy by tunneling [28]. The tunneling step has a time constant τ_{tun} of the same form as (5), but with a different prefactor since interface traps, rather than conduction band states, supply the carriers for tunneling [24]. The two-step process is actually a second-order system described by two time constants; however, since the tunneling and thermal steps occur in series, an effective first-order time constant can be written [24]

$$\tau_{it}(E_t, x) = \tau_{\text{tun}}(E_t, x) + \tau_{\text{SRH}}(E_t - q\mathcal{E}x) \quad (7)$$

where τ_{SRH} is the Shockley–Read–Hall time constant for interface traps [29]:

$$\tau_{\text{SRH}}(E) = \frac{1}{\sigma_n\overline{v}[n_s + n_1(E)] + \sigma_p\overline{v}[p_s + p_1(E)]} \quad (8)$$

with σ_n and σ_p the capture cross sections for electrons and holes (cm^2), $\overline{v}$ the thermal velocity (cm/s), n_s and p_s the surface electron and hole concentrations (cm^{-3}), $n_1(E) =$

$n_s \exp(E_F - E)/kT$, and $p_1(E) = p_s \exp(E - E_F)/kT$. In strong inversion, the thermal time constant for interface traps near the band edges is on the order of 100 ps, so that the process becomes tunneling-limited ($\tau_{it} \approx \tau_{\text{tun}}$). However, for traps near midgap or at low temperatures, the system can become thermally limited ($\tau_{it} \approx \tau_{\text{SRH}}$).

An electron can also tunnel elastically into an excited state of an oxide trap, as shown in Fig 5(c), which subsequently relaxes to its equilibrium ground state with energy loss occurring by cascade phonon emission [30] or lattice relaxation multiphonon emission [31]. The time constant for this phonon-assisted process has the same form as (8), but with a capture cross section that decreases exponentially with depth into the oxide [19], [24]. In addition, the capture cross section for the lattice relaxation process is thermally activated with an energy barrier E_A ranging from 100 to 600 meV [32]. The capture cross section in (8) therefore takes the form $\sigma(x) = \sigma_o \exp(-E_A/kT) \exp[-\lambda(\phi_c, x)]$, where λ is evaluated at the conduction band energy $E = \phi_c$ from which the tunneling transition occurs. Assuming equal capture cross sections for electrons and holes, the time constant for lattice relaxation can therefore be written as shown in (9) at the bottom of this page. Note that the time constants for both inelastic mechanisms (7)–(9) are thermally activated through n_1 and p_1, depend inversely on the surface carrier concentrations, and increase exponentially with the spatial depth of the trap in the oxide. Rigorous deviations of the these time constants from quantum mechanics can be found in [24].

C. Effective Threshold Voltage Shift

The device under test is initially biased in the linear region with drain current

$$I_{DS} = \mu_N \frac{W}{L} |Q_n| V_{DS} \tag{10}$$

where $|Q_n| = C_{\text{ox}}(V_{GS} - V_T - V_{DS}/2)$ is the inversion charge (C/cm^2). By partial differentiation of (10) at constant V_{DS}, the equivalent gate–source voltage shift δV_{GS} required to restore the drain current to its initial value following stress ($\delta I_{DS} = 0$) is

$$\delta V_{GS} = \delta V_T - \frac{|Q_n|\,\delta\mu}{C_{\text{ox}}\,\mu_N}. \tag{11}$$

The contribution to the threshold voltage shift due to a trapped charge density δn_{ot} at energy E_t and depth x is [33]

$$\delta V_T = \frac{q}{C_{\text{ox}}}\left(1 - \frac{x}{t_{\text{ox}}}\right)\delta n_{ot}\delta E_t \delta x. \tag{12}$$

In addition to shifting the threshold voltage, the capture of a carrier by an oxide trap induces a change in the mobility of channel carriers due to increased scattering from the charged trap site. Since the mobility perturbation $\delta\mu$ is correlated with the fluctuation in the number of trapped carriers, it can be written [24], [34]

$$\delta\mu = -\mu_N^2 S(x) q \delta n_{ot} \delta E_t \delta x \approx \frac{-q\mu_N^2 S_o \delta n_{ot} \delta E_t \delta x}{|Q_n|} \tag{13}$$

where $S(x) \approx S_o/|Q_n|$ is the scattering rate (C/V · s) and S_o is a constant with a value of approximately 0.02 V · s/cm^2 for an NMOS transistor [35].

Summarizing these results, traps at energies above the conduction band (cb) or below the valence band (vb) are assumed to communicate with oxide traps by elastic tunneling, while transitions involving traps at energies between the bands must occur either by the interface state assisted (two-step) process (it) or by the lattice relaxation process (lr). Since the bending of the oxide bands shifts trap energy levels with respect to the interface according to (2), traps which fill by one of these mechanisms may empty by another. We shall denote by i/j that component of the threshold voltage shift due to traps which fill by mechanism i during stress and empty by mechanism j following stress, where $i, j \in \{cb, vb, it, lr\}$. Substituting (3), (12), and (13) into (11) and integrating the contributions due to traps at all energies and depths in the oxide, the equivalent gate–source voltage shift due to component i/j is

$$\Delta V_{GS_{(i/j)}}(t_s, t_r) = \frac{q t'_{\text{ox}}}{\epsilon_{\text{ox}}} \int_0^{t'_{ox}} \int_0^{E_G} D_{ot}(E_t, x)\Delta f(E_t, x) \cdot \left(1 - \frac{x}{t'_{ox}}\right)[1 - e^{-t_s/\tau_i(E_t,x)}] \cdot e^{-t_r/\tau_j(E_t,x)} dE_t dx \tag{14}$$

where $E_G \approx 9$ eV is the energy gap of SiO_2 and

$$t'_{\text{ox}} = t_{\text{ox}}(1 + \mu_N S_o). \tag{15}$$

The upper limit of integration in (14) has been changed from t_{ox} to t'_{ox} on the basis that the integrand is negligible at this limit due to the exponential dependence of τ_i on x. Comparison of the form of (14) with (3) and (12) indicates that the equivalent gate–source voltage shift, which is due to both mobility and threshold voltage fluctuations, can be interpreted as an *effective threshold voltage shift* with a scaled oxide thickness given by (15). The total effective threshold voltage shift is just the superposition of all of these components:

$$\Delta V_T(t_s, t_r) = \sum_{i,j} \Delta V_{GS_{(i/j)}}(t_s, t_r). \tag{16}$$

D. Approximate Voltage and Time Dependence

Insight into the voltage and time dependence of (14) can be obtained by consideration of the important special case of strong inversion operation at room temperature. In order to obtain closed-form solutions, we make the following assumptions.

$$\tau_{lr}(E_t, x) = \frac{1}{\sigma_o \overline{v}[n_s + n_1(E_t - q\mathcal{E}x) + p_s + p_1(E_t - q\mathcal{E}x)]} e^{E_A/kT} e^{\lambda(\phi_c, x)}. \tag{9}$$

1) Uniform oxide trap distribution $D_{ot}(E_t, x) = D_{ot}$.
2) Step approximation to the Fermi function:

$$f(E, E_F) = \begin{cases} 1 & E \geq E_F \\ 0 & E < E_F. \end{cases}$$

3) Negligible mobility fluctuations ($\mu_N S_o = 0$).
4) $x \ll t_{\text{ox}}$ since typical tunneling depths are less than 20 Å while 250 Å $\leq t_{\text{ox}} \leq$ 750 Å.
5) Rectangular tunneling barrier: Using assumption 4), this approximation results in less than 10% error in average barrier height for 5 V across a 500 Å oxide.
6) Pinning of the Fermi level at the conduction band edge during positive stress ($E_F \approx E_c$).
7) $V_{BS} = 0$.

Under these assumptions, all three processes 5), 7), and 9) are tunneling-limited with a time constant of the form $\tau_s(x) = \tau_{os} \exp(2Kx)$ during stress and $\tau_r(x) = \tau_{or} \exp(2Kx)$ during relaxation, where the prefactors generally differ due to the bias dependence of (8) and (9). As shown in the Appendix, the total threshold voltage shift can then be written as the sum of two components:

$$\Delta V_T = \Delta V_\psi + \Delta V_\mathcal{E} \tag{17}$$

where

$$\Delta V_\psi = \frac{q^2 D_{ot} t_{\text{ox}}}{2K\epsilon_{\text{ox}}} \Delta\psi_s \ln\left(1 + \frac{\tau_{or} t_s}{\tau_{os} t_r}\right) \tag{18}$$

$$\Delta V_\mathcal{E} = \frac{q^2 D_{ot}}{4K^2 \epsilon_{ox}} \Delta V_{\text{ox}} \ln\left(1 + \frac{\tau_{or} t_s}{\tau_{os} t_r}\right) \cdot \left[\gamma + \frac{1}{2}\ln\left(1 + \frac{\tau_{or} t_s}{\tau_{os} t_r}\right) + \ln\left(\frac{t_r}{\tau_{or}}\right)\right]. \tag{19}$$

As shown in Fig. 4, the components ΔV_ψ and $\Delta V_\mathcal{E}$ correspond to windows in E–x space containing those traps that change occupancy during the measurement sequence, thereby contributing a nonzero value to the integrand in (14). The *surface potential component* ΔV_ψ is due to traps at energies within the silicon bandgap that change occupancy as a consequence of the change in Fermi level or, equivalently, the change in surface potential $\Delta\psi_s = (E_{F0} - E_F)/q$ that occurs during stress. On a microscopic level, the traps that comprise this component therefore fill and empty via the it/it or lr/lr mechanisms described above. The *band-bending component* $\Delta V_\mathcal{E}$, which is driven by the change in voltage $\Delta V_{ox} = (\mathcal{E}_o - \mathcal{E})t_{ox}$ across the oxide, is due to traps that are lowered below the conduction band during stress and raised above it when the stress is removed. The traps which contribute to this component therefore fill and empty by the lr/cb and/or it/cb processes.

Equations (18) and (19) show that a pure logarithmic time dependence, as predicted by previous models [7], [23], results only when band bending in the oxide is neglected ($\Delta V_{ox} = 0$) and only in the limits $t_s \gg t_r$, both assumptions of which are violated under typical circuit operating conditions. For $t_s \gg t_r$, the inclusion of the band-bending term contributes a component that goes as $\log^2$(time) to the usual log(time) dependence, while for $t_r \gg t_s$, the present theory predicts that the decay will approach $1/t_r$.

IV. Results

While (18) and (19) describe the general features of the threshold voltage shifts with voltage and time, the accurate modeling of their temperature and bias dependence demands a more exact treatment. In this section, (14) is integrated numerically using the energy-dependent time constants of (5), (7), and (9), with a trapezoidal potential barrier and Fermi–Dirac statistics. All of the parameters in (14) are known quantities, with the exception of $D_{ot}(E_t, x)$, which is treated as an adjustable fitting parameter. The trap distribution can be modeled quite generally by a function of the form

$$D_{ot}(E_t, x) = D_0 e^{-x/x_o} + \sum_{i=1}^{N} D_i g_i(E_t)[1 - g_i(E_t)]e^{-x/x_i} \tag{20}$$

where

$$g_i(E_t) = \frac{1}{1 + \exp[(E_i - E_t)/\Delta E_i]}.$$

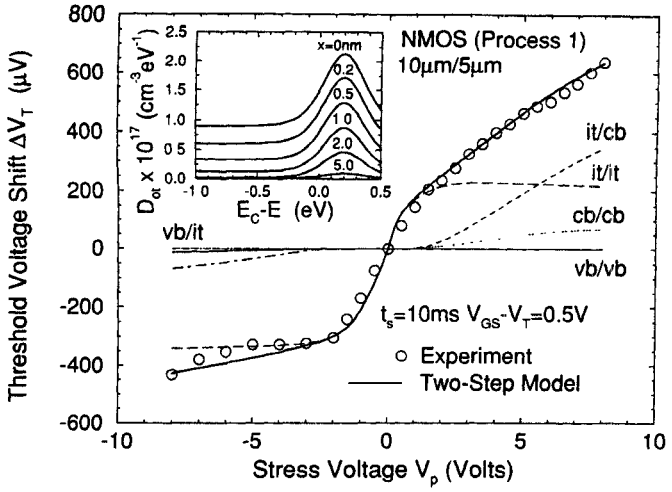
Equation (20) is the superposition of a uniform trap density D_0 and N peaks at the energies E_i. By varying the number and the spread ΔE_i of these peaks, a wide variety of trap energy distributions can be modeled ranging from uniform to discrete levels. The trap density is assumed to decay with increasing depth into the oxide with a characteristic length x_i. Since only traps within 20 Å of the interface are of interest, this assumption is not restrictive and simplifies the numerical integration.

In the following sections, transient threshold voltage shifts are characterized over a wide range of bias and stress conditions and compared with simulations based on this model. Without loss of generality, an NMOS transistor is assumed. Differences between threshold voltage shifts in NMOS and PMOS transistors are addressed in Section V-B.

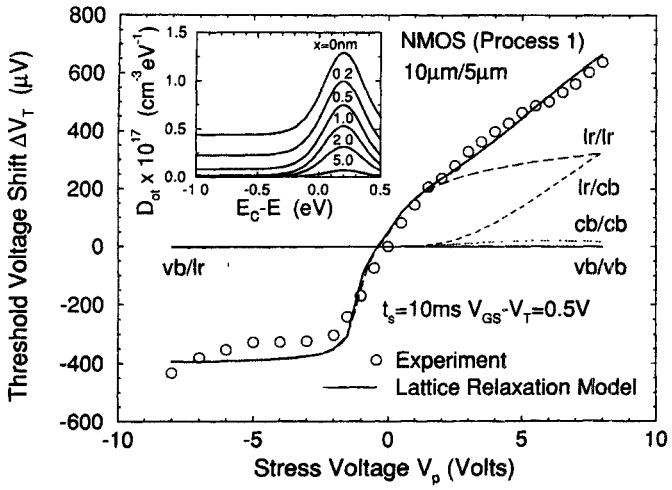
A. Stress Voltage

Fig. 6(a) plots the measured threshold voltage shift versus applied stress voltage for a stress time of 10 ms. Simulations using the two-step model (7) are also shown. The trap distribution in (20) was fit to the data with $N = 1$ and the parameter values in Table II. As shown in the inset to Fig. 6(a), this trap distribution increases with energy above the conduction band, peaking at about 0.2 eV above the band edge, consistent with previous findings [34], [36], [37]. Also shown in Table II are values for the time constant prefactors $\tau_{cb0}(\tau_{it0})$ and $\tau_{cb1}(\tau_{it1})$ for tunneling to and from conduction band states (interface traps) during relaxation and stress, respectively. In Fig. 6(b), the same data are shown, together with simulations using the lattice relaxation model (9). While the forms of the extracted trap densities in Fig. 6(a) and (b) are identical, their magnitudes differ somewhat depending on which model is used. The voltage dependence can be explained equally well by either mechanism, making it difficult to experimentally resolve their relative contributions.

For positive stress, most of the applied voltage is dropped across the oxide ($\Delta V_{\text{ox}} \approx V_p$), and the it/cb and lr/cb components, which comprise the band-bending term $\Delta V_\mathcal{E}$,



(a)



(b)

Fig. 6. Measured and simulated dependence of threshold voltage shift on stress voltage. (a) Two-step model. (b) Lattice relaxation model.

TABLE II
PARAMETER VALUES USED IN SIMULATIONS (TWO-STEP MODEL)

Model Parameter	Units	Process 1	Process 2	Process 5
D_0	$cm^{-3}\,eV^{-1}$	7.0×10^{16}	6.0×10^{16}	2.5×10^{16}
D_1	$cm^{-3}\,eV^{-1}$	2.3×10^{17}	2.9×10^{17}	2.5×10^{17}
E_1	eV	2.9	3.0	3.1
ΔE_1	eV	0.1	0.1	0.03
x_0	Å	5.0	5.0	5.0
x_1	Å	20.0	20.0	50.0
τ_{cb0}	ns	0.1	9.0	0.2
τ_{cb1}	ns	0.1	9.0	1.0
τ_{it0}	ns	0.1	9.0	8.0
τ_{it1}	ns	0.1	9.0	1.0
$\sigma_n, \sigma_p, \sigma_o$	cm^2	1.0×10^{-15}	1.0×10^{-15}	1.0×10^{-15}

increase nearly linearly with voltage, as predicted by (19). The components it/it and lr/lr follow the voltage dependence of ΔV_ψ in (18), saturating in strong inversion and accumulation when the surface potential becomes pinned ($\Delta\psi_s \approx 0$). Note, however, that the lr/lr component does not completely clamp

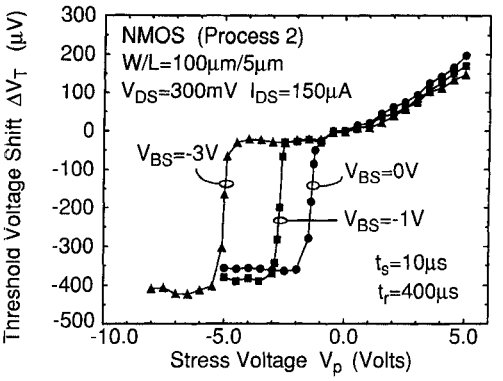


Fig. 7. Threshold voltage shift versus stress voltage with substrate bias as a parameter.

because the prefactor of the time constant in (9) is voltage dependent. In strong inversion, the prefactor decreases with increasing gate voltage as $\tau_o \approx (\sigma_n \bar{v} n_s)^{-1} \propto [C_{ox}(V_{GS} - V_T)]^{-1}$. This leads to a slight increase in ΔV_T over that predicted from the change in surface potential alone.

For negative stress voltages which pulse the surface into depletion, the behavior of ΔV_T with voltage is strongly dependent on temperature and stress time. When the stress time is long compared with the time constants for oxide traps within the bandgap ($t_s \gg \tau_{it}, \tau_{lr}$), electron emission can follow the applied voltage, and the ΔV_ψ components it/it and lr/lr track the change in surface potential as the Fermi level sweeps down through the bandgap, as shown in Fig. 6. At shorter stress times or lower temperatures, the time constants in (7) and (9) can become too long for traps within the bandgap to follow the applied voltage ($t_s \ll \tau_{it}, \tau_{lr}$). Under these conditions, electrons remain trapped until the surface accumulates, at which point the time constants become sufficiently short for emission to occur, producing an abrupt negative threshold voltage shift as shown in Fig. 7.

After accumulation is reached, further increases in the magnitude of V_p result in negligible changes in threshold voltage shift for two reasons. First, the tunneling probability for holes is much smaller than for electrons, as explained above. Second, it is well known that traps above midgap tend to be acceptor-like, while those below are donor-like [38]. The discussion above assumed acceptor-like traps, which empty and become neutral during negative stress. However, large negative stress voltages drive the Fermi level down to the valence band edge, causing donor-like traps to empty and become positively charged. When the surface returns to inversion, the high electric field (10^6–10^8 V/cm) created across the thin oxide barrier by the dipole layer of positively charged donor traps and negative inversion layer charge causes electrons to be recaptured in a period too short to be measured by the circuit. This high field condition does not occur for positive stress because occupied acceptor traps and inversion layer electrons have the same polarity.

B. Substrate Bias

Fig. 7 shows the effect of a substrate–source voltage V_{BS} on the $\Delta V_T - V_p$ characteristics. As V_{BS} was varied, the initial

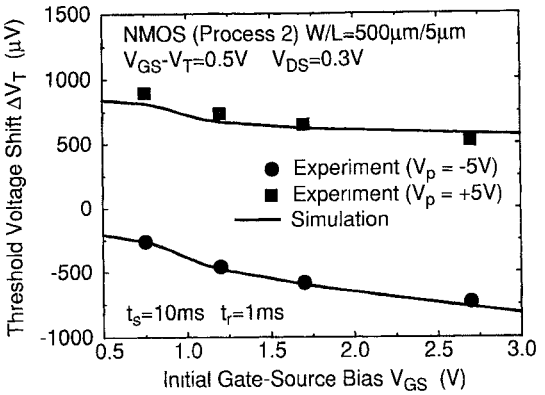


Fig. 8. Dependence of threshold voltage shift on prestress gate–source voltage bias.

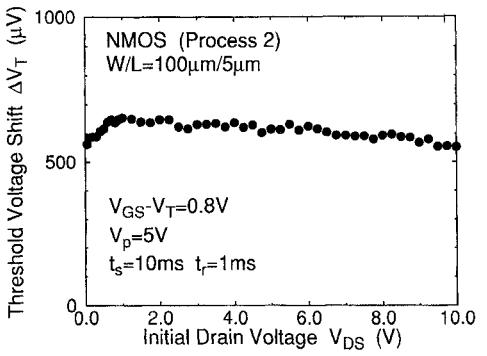


Fig. 9. Drain bias dependence.

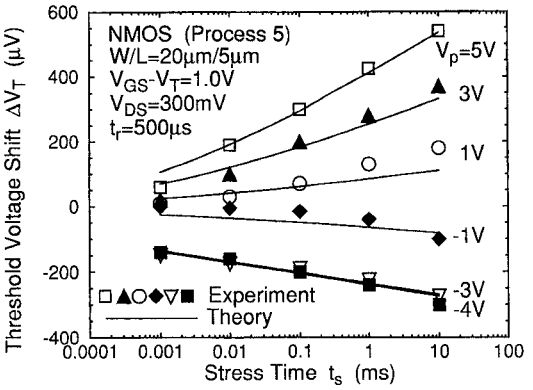


Fig. 10. Threshold voltage shift versus stress time with stress voltage as a parameter.

gate–source bias was adjusted to compensate for the back-gate effect, maintaining the prestress value of the drain current at a constant level. The characteristics therefore change negligibly in the positive direction because the relationship between the electron quasi-Fermi level and the conduction band edge is fixed. However, since the application of a substrate bias V_{BS} shifts the point at which accumulation occurs to a more negative voltage, the steps in the characteristics shift along the V_p axis by an amount [24]

$$\Delta V_p = V_{BS} - \gamma[\sqrt{\phi_s - V_{BS}} - \sqrt{\phi_s}] \tag{21}$$

where ϕ_s is the surface potential in strong inversion and $\gamma = \sqrt{2q\epsilon_s N_A}/C_{ox}$ is the back-gate effect parameter. The first term in (21) accounts for the splitting of the electron and hole quasi-Fermi levels at the interface by the applied substrate bias. The second term is just the voltage required to cancel the increase in gate–source bias which was applied to compensate for the back-gate effect.

C. Initial Gate–Source Bias

Previously, it had been reported that the threshold voltage shifts produced in NMOS transistors under positive gate–source voltage stress decrease with increasing drain current [5]. This observation can now be explained by the tunneling model. Assuming that the Fermi level becomes pinned at the conduction and valence band edges in strong inversion and accumulation, respectively, the maximum amplitude of the surface potential component ΔV_ψ is proportional to $|E_c - E_{F0}|$ for positive and to $|E_{F0} - E_v|$ for negative stress. As the initial level of inversion increases with increasing gate bias, E_{F0} moves closer to E_c and farther from E_v. This causes the magnitude of the threshold voltage shift to decrease for positive and increase for negative stress, as shown in Fig. 8.

D. Drain–Source Voltage

While measurements at low drain voltages enable the tunneling mechanism to be isolated from competing effects, MOSFET's in analog circuits are normally biased in saturation where the transconductance and output resistance are high. At arbitrary drain biases, (1) no longer holds, and a change in power dissipation will generally occur during stress. Fig. 9 shows the initial threshold voltage shift plotted as a function of the prestress drain–source bias. The threshold voltage shift is nearly independent of drain voltage, consistent with the tunneling model. Since hot carriers and self-heating effects, if present, would tend to be exacerbated at higher drain voltages, it can be concluded that these mechanisms are negligible under the pulsed stress conditions used in this study, and that direct tunneling dominates the threshold voltage shifts well into saturation.

E. Stress and Relaxation Time

Fig. 10 shows the dependence of the initial threshold voltage shift on stress time, with stress voltage as a parameter. Since the measurement delay t_r is much less than the stress time t_s, the contribution due to ΔV_ψ in (18) is approximately logarithmic. Some curvature is apparent, however, due to the log squared term in $\Delta V_\mathcal{E}$ (19). The linear voltage dependence for positive stress and the sharp nonlinearity for negative stress can also be seen.

The relaxation of the threshold voltage shift as a function of time following stress is shown in Fig. 11. Deviations from a strict logarithmic decay, which cannot be explained by previous theories [7], [23], are predicted by the new model when the effects of band bending in the oxide, Fermi–Dirac statistics, and nonuniform oxide trap distributions are taken into account.

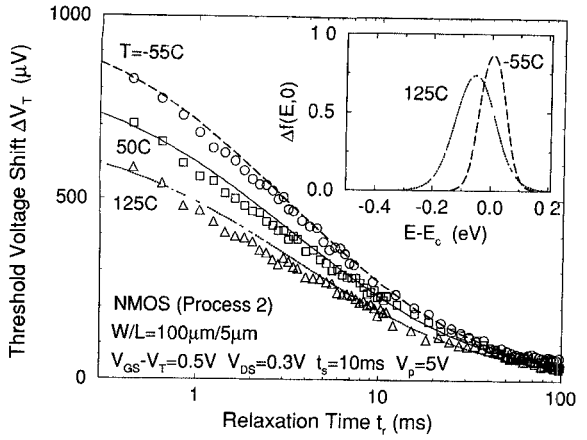


Fig. 11. Measured and simulated threshold voltage relaxation transients at three different temperatures. Inset: the function $\Delta f(E, x)$ at $x = 0$.

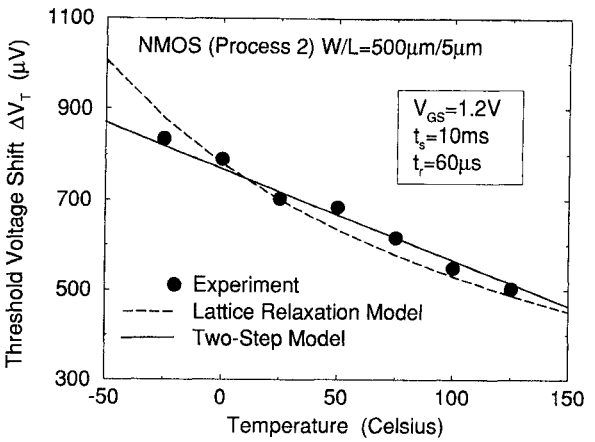


Fig. 12. Initial threshold voltge shift versus temperature.

F. Temperature

Threshold voltage shifts were measured over temperature by placing the integrated circuit test chip (Fig. 1) in an oven or refrigerator with the rest of the test circuitry maintained at room temperature. The tunneling model contains three primary sources of temperature dependence.

1) With increasing temperature, the width of the Fermi function spreads and the Fermi level moves lower in the bandgap. The function $\Delta f(E_t, x)$ in (4), which determines the range of trap energies that change occupancy during the measurement sequence, is plotted in the inset to Fig. 11 versus energy above the conduction band at $x = 0$. With increasing temperature, Δf increases at some energies and decreases for others. The temperature coefficient of the threshold voltage shift therefore depends on the energy and spatial distribution of oxide traps. The graph shows that, for traps concentrated above the conduction band as shown in Fig. 6, the temperature coefficient will be negative, as observed. This mechanism alone can explain the temperature dependence for positive stress voltages, as shown in Fig. 11.

2) The time constants τ_{it} and τ_{lr} of (7) and (9), which contribute mainly for negative gate voltage pulses, are thermally activated and decrease strongly with increasing temperature.

3) The $T^{-3/2}$ temperature dependence of mobility μ_N in (15) leads to a decreasing ΔV_T with increasing temperature due to correlated mobility fluctuations.

Fig. 12 plots the initial threshold voltage shift versus temperature, together with simulations using the two-step model and the lattice relaxation model with correlated mobility fluctuations. A uniform distribution of activation energies E_A was assumed, with $S_o = 0.01$ V·s/cm^2 and $\mu_N = 500$ cm^2/V·s. The lattice relaxation model leads to an increasing threshold voltage shift with temperature due to the activated time constant (9). However, this increase is offset by the negative temperature coefficient of the correlated mobility term. The data fall somewhere between the two simulated curves, and can be explained equally well by either inelastic mechanism or by a linear combination of the two. Experimentally, the temperature coefficient for positive stress varies from approximately -0.1 to $-3.25\,\mu$V/°C.

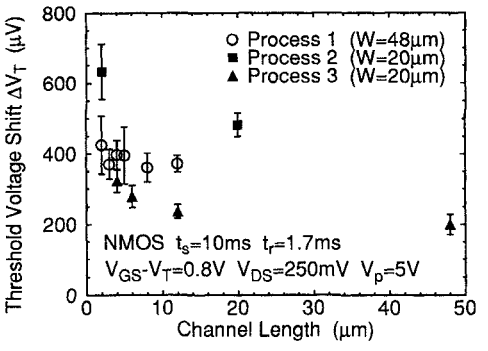
G. Channel Width and Length

The threshold voltage shift is observed to decrease slightly with increasing channel length and width, as shown in Fig. 13. This can be explained as an edge effect, due to higher trap densities around the periphery of the gate oxide arising from damage created during implantation of the source–drain regions or from higher trap densities in the field oxide that forms the bird's beak [39]. Consider a uniform trap density D_{ot} in the interior region of the gate oxide, as shown in the inset of Fig. 14, and strips of width ΔW along the bird's beak and ΔL bordering on the source–drain regions in which the trap density has enhanced values D_{ot}^W and D_{ot}^L, respectively. The threshold voltage shift is proportional to the total charge divided by the total oxide capacitance, integrated over all energies and depths. Since the above theory assumed a trap density which was uniform over area, (14) must be modified by the substitution shown at the bottom of this page. For $\Delta W \ll W$ and $\Delta L \ll L$, which is valid for the devices used in this study, ΔV_T as calculated in (14) must therefore be multiplied by a geometry-dependent prefactor:

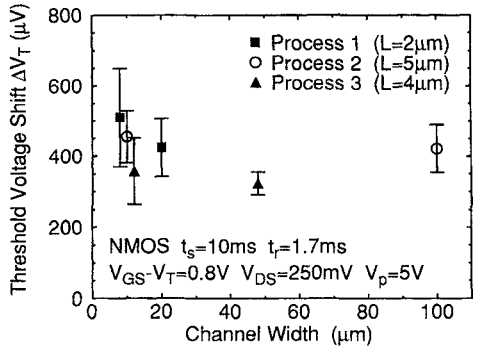
$$\Delta V_T'(L, W) = \left[1 + \frac{\gamma_l}{L} + \frac{\gamma_w}{W}\right]\Delta V_T \tag{22}$$

where $\gamma_l = 2\Delta L D_{ot}^L / D_{ot}, \gamma_w = 2\Delta W D_{ot}^W / D_{ot}$, and ΔV_T is the asymptotic value of the threshold voltage shift for large L and W. Simulations using (22) with $\gamma_l = 1.24\,\mu$m are plotted in Fig. 14, and show excellent agreement with the data.

$$\frac{D_{ot}WL}{C_{\rm ox}WL} \rightarrow \frac{D_{ot}(W - 2\Delta W)(L - 2\Delta L) + 2W\Delta L D_{ot}^L + 2\Delta W(L - 2\Delta L)D_{ot}^W}{C_{\rm ox}WL}$$



(a)



(b)

Fig. 13. Dependence of initial threshold voltage shift on (a) channel length, (b) channel width. Error bars indicate ±1 one standard deviation.

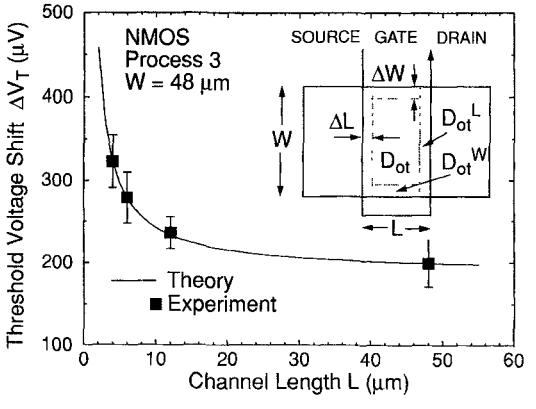


Fig. 14. Measured and simulated channel length dependence.

V. Discussion

A. Scaling

The behavior of transient threshold voltage shifts as device feature sizes are scaled is of interest for future process generations. The surface potential component ΔV_ψ of the threshold voltage shift in (18) is proportional to t_{ox} and decreases as oxide thickness is reduced. However, the band-bending term $\Delta V_\mathcal{E}$ in (19), which dominates in strong inversion, decreases with decreasing oxide thickness only if voltage levels scale proportionately; otherwise, it remains constant. For arbitrary bias conditions, where the threshold voltage shift (17) is a superposition of these two components, the tunneling model therefore predicts that the threshold voltage shift will decrease as oxide thickness is reduced, but at a sublinear rate. Further experimental work is needed to confirm these results.

B. Differences Between NMOS and PMOS

Table I indicates that, under positive stress, threshold voltage shifts in NMOS transistors are about an order of magnitude larger than in PMOS. Threshold voltage shifts for negative stress voltages are typically much smaller than for positive stress in both channel types. The general features of these observations are consistent with previously reported results [5], although their absolute magnitudes differ somewhat due to variations in trap energy distributions between processes. These results can be explained on the basis of two factors: 1) the extracted trap densities, as depicted in Fig. 6, consist primarily of acceptor-like traps located energetically above the conduction band, and 2) the tunneling probability for holes is much smaller than for electrons.

For an NMOS transistor in strong inversion, the Fermi level at the surface is initially close to the conduction band, and a threshold voltage shift proportional to stress voltage occurs as acceptor traps are lowered below the Fermi level under positive stress. For traps above the conduction band as indicated, the change in trap occupancy during positive stress will be smaller for PMOS transistors because the Fermi level is initially close to the valence band and must traverse the entire bandgap before trap occupancies can change significantly. Furthermore, because the traps are acceptor-like, any traps which capture electrons during stress become negatively charged and discharge rapidly by field emission when the surface returns to inversion and becomes flooded with holes. Donor-like traps in the lower half of the bandgap can cause a threshold voltage shift in PMOS transistors under positive stress, but because the time constants for these traps depend inversely on carrier concentration according to (8) and (9), this shift is not observed until the surface accumulates.

For negative stress voltage pulses, the Fermi level in PMOS is pinned near the valence band, and threshold voltage shifts are negligible due to lower trap densities at these energies and to the small probability of hole tunneling. NMOS transistors exhibit some shift in the negative direction due to the detrapping of electrons for acceptor-like traps as the Fermi level sweeps down through the bandgap. Once the Fermi level becomes pinned at the valence band, however, no further increase in threshold voltage shift occurs due to negligible hole tunneling and to the rapid recapture of electrons by empty donor traps when the surface returns to inversion.

In addition to the above considerations, several additional factors contribute to lower threshold voltage shifts in PMOS than in NMOS transistors: 1) PMOS devices tend to have buried channels, in which a potential barrier physically separates inversion layer holes from the Si–SiO_2 interface, thereby reducing trapping; 2) the mobility fluctuation term $\mu_N S_o$ in (15) is generally about an order of magnitude smaller in PMOS than in NMOS transistors [35]; and 3) differences in trap densities between NMOS and PMOS devices arising from processing differences may also account for some of the discrepency.

C. Correlation with 1/f Noise

In a model of $1/f$ noise based on carrier number fluctuations, the power spectral density of the equivalent gate voltage noise is [20], [34]

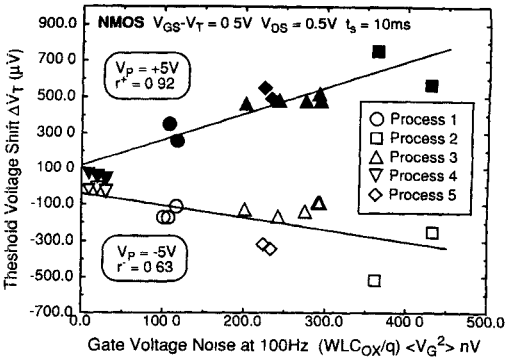
$$\langle V_{GS}^2 \rangle = \frac{q^2(1+\mu_N S_o)}{WLC_{\text{ox}}^2} \int_0^{t_{\text{ox}}} dx \int_{E_v}^{E_c} \cdot dED_{ot}(E,x) f(E)[1-f(E)] \frac{\tau(x)}{1+\omega^2\tau^2(x)}$$

with $\tau(x) = \tau_o \exp(2Kx)$. This equation is similar in form to (14), with the important distinction that the factor $f(E)[1-f(E)]$ resembles an impulse centered around the electron quasi-Fermi level E_{Fn}, so that $1/f$ noise probes the oxide trap density in the vicinity of single energy, whereas the function $\Delta f(E,x)$ in (4) probes the trap density over a much wider range of energies, as shown in Fig. 11.

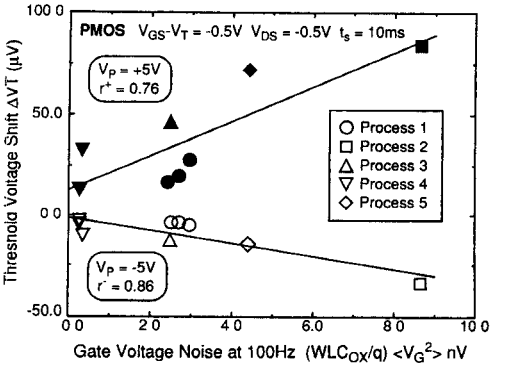
In Fig. 15, threshold voltage shifts following stress at ± 5 V for 10 ms are plotted versus the measured gate voltage noise in each of the five processes. The equivalent gate voltage noise $\langle V_{GS}^2 \rangle$ is measured at a frequency of 100 Hz, where the $1/f$ component dominates. Also shown are the linear regression lines and correlation coefficients r^+ and r^- between $1/f$ noise and threshold voltage shifts in the positive and negative directions, respectively. A strong correlation is observed in the case of positive (negative) stress pulses in NMOS (PMOS) because the Fermi level is pinned near the conduction band (valence band) edge throughout the measurement. The trap densities probed by the $1/f$ and ΔV_T measurements are therefore nearly identical, i.e., the function Δf reduces to $f(1-f)$. The correlation is weaker for pulses of the opposite polarity because the Fermi level traverses the bandgap during large-signal stress and the threshold voltage measurement is affected by traps far outside the range which contribute to $1/f$ noise. These results support the conclusion that $1/f$ noise and transient threshold voltage shifts are simply small- and large-signal manifestations, respectively, of the same underlying mechanism.

D. Circuit Model

The accurate model presented above is valuable for understanding the origin and behavior of transient threshold voltage shifts based on the underlying physics. However, for evaluating their impact on circuit performance during the design process, a simple and computationally efficient model is required. Fig. 16(a) shows a simple lumped element subcircuit model which can be used to evaluate threshold voltage relaxation effects in a circuit simulation program such as SPICE. The gate–source voltage across the device of interest is replicated by a voltage-controlled voltage source which drives an RC ladder network similar to that used to model dielectric relaxation in capacitors [40], [41]. The currents into the rungs of this ladder model charging currents into traps at different depths in the oxide. These currents are summed, multiplied by a process-dependent gain factor (A), and integrated onto the oxide capacitance (C_{OX}). The capacitor voltage, which represents the threshold voltage shift, is inserted in series with the gate of the device of interest via a second voltage-controlled voltage source.



(a)



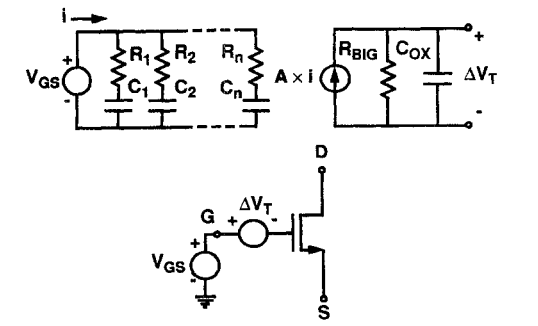
(b)

Fig. 15. Threshold voltage shift versus $1/f$ noise in five different processes. (a) NMOS. (b) PMOS.

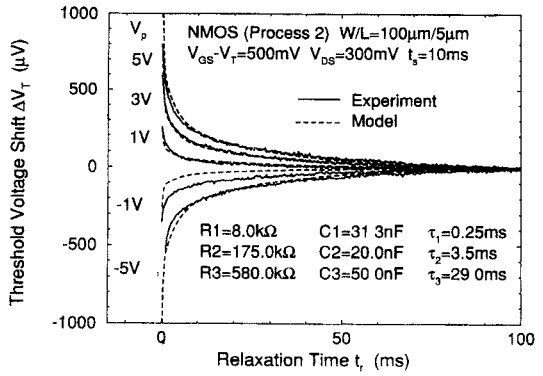
The values of the parameters in this model can be obtained by matching the response of the circuit to measured relaxation transients. In Fig. 16(b), only three time constants were used and provide a good fit to the data for positive stress. For negative stress voltages, the fit is less satisfactory because the nonlinearity of the threshold voltage shift cannot be modeled by a linear circuit. This nonlinearity, which arises from the MOS surface potential gate–voltage relationship and the voltage dependence of the time constants, can be modeled by making the gain (A) a nonlinear function of stress voltage. In practice, however, accurate modeling of threshold voltage shifts for accumulating stress pulses is less important because they can be suppressed using the methods described below.

E. Minimization of Threshold Voltage Shifts in Circuits

Based on the results of this study, several techniques for the minimization of threshold voltage shifts in circuits can be recommended. From a processing standpoint, the reduction of oxide trap densities is crucial. For example, wet gate oxides are known to contain higher trap densities than dry oxides, and should be avoided. Buried channels offer the possibility of reducing trapping by separating inversion layer charge carriers from oxide traps. In general, the preceding results indicate that a process designed for low $1/f$ noise

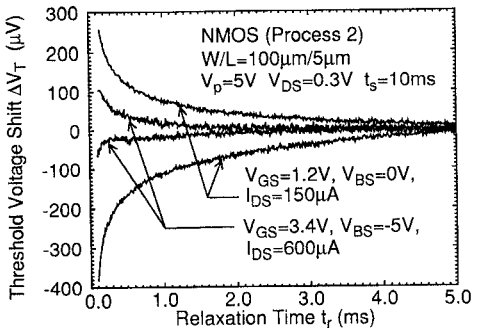


(a)

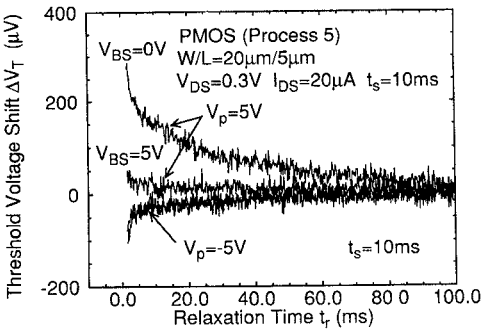


(b)

Fig. 16. Circuit model. (a) Schematic. (b) Measured and SPICE simulated threshold voltage relaxation transients.



(a)



(b)

Fig. 17. Bias conditions for the minimization of threshold voltage shifts. (a) NMOS. (b) PMOS.

will also show good immunity to large-signal charge trapping effects.

From a circuit perspective, MOS transistors, particularly on input stages, should be protected from large-signal gate–source voltage transients. When this is not possible, PMOS transistors offer superior immunity to transient threshold voltage shifts, and should be used in preference to NMOS. Bias conditions can also be chosen to densensitize a device to transient gate-source voltage stress. Since the threshold voltage shift for negative (positive) pulses in NMOS (PMOS) transistors are small until accumulation is reached (Fig. 7), threshold voltage shifts in this direction can be suppressed by tying the back gate to the negative (positive) supply voltage. The data in Fig. 8 suggest that strong inversion operation can be used to suppress threshold voltage shifts for positive (negative) gate–source voltage pulses in NMOS (PMOS) transistors. The results of applying both of these techniques to an NMOS transistor are shown in Fig. 17(a). The shift in the negative direction is almost entirely eliminated by the application of a substrate bias, while the positive shift is reduced by about 50% by the increase in gate bias. These techniques are even more effective when used on PMOS transistors, in which negative polarity threshold voltage shifts are inherently small (Table I). The shift in the positive direction can be suppressed by connecting the back gate to the positive voltage supply, as shown in Fig. 17(b). This configuration is nearly free from threshold voltage shifts for either polarity of stress, and can be used to advantage in sensitive areas such as the input stages of opamps and comparators.

VI. Conclusions

With the expanding applications of CMOS and BiCMOS technologies for the implementation of analog integrated circuits, transient threshold voltage shifts will impose increasingly important limitations on speed and accuracy. While this problem can degrade circuit accuracy at the 12 b level and above, it can be successfully minimized and modeled using the techniques described in this paper.

Measurements and theory have been presented which demonstrate that, under large-signal transient stress conditions, threshold voltage shifts are caused by the direct tunnel exchange of channel charge carriers with preexisting traps in the oxide, within 20Å of the Si-SiO_2 interface. In addition to elastic tunneling transitions involving traps at the conduction and valence band energies, the voltage, time and temperature dependence of the threshold voltage shifts indicate that inelastic transitions occur into oxide traps at energies within the silicon bandgap. Two microscopic mechanisms were proposed for this charge exchange, in which tunneling occurs in conjunction with lattice relaxation in the oxide or through the assistance of interface traps. Both mechanisms are capable of explaining the experimental results and, from a macroscopic modeling point of view, are indistinguishable. However, the use of this technique for the accurate extraction of oxide trap densities must await the development of more precise experimental methods capable of resolving the relative contributions of these two inelastic components. In addition to identifying the physical mechanisms responsible for transient threshold voltage shifts in MOSFET's, this

work has applications to explaining drifts in II-VI and II-V MISFET's [42], [43] and to modeling the effects of oxide traps on other large-signal measurements such as deep-level transient spectroscopy (DLTS) [12], [25] and charge pumping [44].

APPENDIX

Assumption 2) of Section III-D implies that

$$f(E_t - q\mathcal{E}x, E_F) - f(E_t - q\mathcal{E}_o x, E_{F0}) = \begin{cases} +1 & E_F + q\mathcal{E}x \le E_t < E_{F0} + q\mathcal{E}_o x \quad (V_p > 0) \\ -1 & E_{F0} + q\mathcal{E}_o x \le E_t < E_F + q\mathcal{E}x \quad (V_p < 0) \\ 0 & \text{otherwise.} \end{cases}$$

Substituting this result into (14) and invoking the remaining assumptions of Section III-D, the threshold voltage shift is

$$\Delta V_T(t_s, t_r) = \pm \frac{qD_{ot}}{C_{\text{ox}}} \int_0^{t_{\text{ox}}} \int_{E_F + q\mathcal{E}x}^{E_{F0} + q\mathcal{E}_o x} \cdot [1 - e^{-t_s/\tau_s(x)}] e^{-t_r/\tau_r(x)} dE_t dx$$

where the plus sign applies for $V_p > 0$ and the minus sign for $V_p < 0$. The energy integration can be evaluated trivially to yield two terms:

$$\Delta V_\psi = \pm \frac{qD_{ot}}{C_{\text{ox}}} (E_{F0} - E_F) \int_0^{t_{\text{ox}}} \cdot [1 - e^{-t_s/\tau_s(x)}] e^{-t_r/\tau_r(x)} dx \tag{23}$$

$$\Delta V_\mathcal{E} = \pm \frac{q^2 D_{ot}}{C_{\text{ox}}} (\mathcal{E}_o - \mathcal{E}) \int_0^{t_{\text{ox}}} x[1 - e^{-t_s/\tau_s}] e^{-t_r/\tau_r} dx. \tag{24}$$

The integral in (23) can be carried out by changing the variable of integration from x to $y = t/\tau_o \exp(-2Kx)$ and recognizing that

$$\int_0^{t_{\text{ox}}} \exp\left[-\frac{t}{\tau_o}\exp(-2Kx)\right] dx = \frac{1}{2K} \int_{t/\tau_1}^{t/\tau_o} \frac{e^{-y}}{y} dy = \frac{1}{2K}\left[E_1\left(\frac{t}{\tau_1}\right) - E_1\left(\frac{t}{\tau_o}\right)\right]$$

where $\tau_1 = \tau_o \exp(2Kt_{\text{ox}})$ and $E_1(z)$ is the exponential integral function [45]

$$E_1(z) = \int_z^\infty \frac{e^{-x}}{x} dx$$

which has the properties that $\lim_{z\to\infty} E_1(z) = 0$ and $E_1(z) \approx -\gamma - \ln z$ for small z, where $\gamma = 0.5772$ is Euler's constant. Thus,

$$\Delta V_\psi = \pm \frac{qD_{ot}}{2KC_{\text{ox}}} (E_{F0} - E_F)\left\{E_1\left(\frac{t_r}{\tau_{1r}}\right) - E_1\left(\frac{t_r}{\tau_{or}}\right) + E_1\left(\frac{t_s}{\tau_{1s}} + \frac{t_r}{\tau_{1r}}\right) + E_1\left(\frac{t_s}{\tau_{os}} - \frac{t_r}{\tau_{or}}\right)\right\}. \tag{25}$$

Since τ_{os} and τ_{or} are on the order of 10^{-10} s, it is always the case in our measurements that $t_s \gg \tau_{os}$ and $t_r \gg \tau_{or}$ so that the second and third exponential integrals in (25) can be neglected. Since $t_{\text{ox}} > 250$ Å while the maximum tunneling depth is only 10–20 Å, the time constants $\tau_{1s} = \tau_{os}\exp(2Kt_{\text{ox}})$ and $\tau_{1r} = \tau_{or}\exp(2Kt_{\text{ox}})$ are very large relative to time scales of the measurement, so that, using the series expansion of the exponential integral, (25) can be rearranged to give (18).

The integral in (24) can be readily evaluated by noting that

$$-\frac{d}{dt}\int_0^{t_{\text{ox}}} xe^{-t/\tau} dx = \int_0^{t_{\text{ox}}} xe^{-t/\tau} \frac{dx}{\tau}.$$

Using the substitution $u = t/\tau$,

$$\int_0^{t_{\text{ox}}} xe^{-t/\tau} \frac{dx}{\tau} = \frac{1}{4K^2 t} \int_{t/\tau_1}^{t/\tau_o} \left[\ln\left(\frac{t}{\tau_o}\right) - \ln(u)\right] e^{-u} du = \frac{1}{4K^2 t}\left[\ln\left(\frac{t}{\tau_o}\right) + \gamma\right] \tag{26}$$

where the lower and upper integration limits in (26) have been replaced by zero and infinity, respectively, since $\tau_o \ll t \ll \tau_1$ for the time scales of interest. Thus,

$$\int_0^{t_{\text{ox}}} xe^{-t/\tau} dx = -\frac{1}{4K^2}\left[\frac{1}{2}\ln^2\left(\frac{t}{\tau_o}\right) + \gamma \ln\left(\frac{t}{\tau_o}\right)\right]$$

so that $\Delta V_\mathcal{E}$ is

$$\Delta V_\mathcal{E} = \frac{q^2 D_{ot}}{4K^2 \epsilon_{\text{ox}}} \Delta V_{\text{ox}} \left\{\frac{1}{2}\left[\ln^2\left(\frac{t_s}{\tau_{os}} + \frac{t_r}{\tau_{or}}\right) - \ln^2\left(\frac{t_r}{\tau_{or}}\right)\right] + \gamma\left[\ln\left(\frac{t_s}{\tau_{os}} + \frac{t_r}{\tau_{or}}\right) - \ln\left(\frac{t_r}{\tau_{or}}\right)\right]\right\} \tag{27}$$

which can be rewritten, by simple algebraic manipulations, as (19).

ACKNOWLEDGMENT

The authors wish to thank Analog Devices Inc. and the M.I.T. Microsystems Technology Laboratories for fabrication of the devices used in this study. They especially thank Prof. S. D. Senturia and Prof. C. G. Sodini at M.I.T. for technical assistance and numerous helpful discussions.

REFERENCES

[1] I. A. Young and D. A. Hodges, "MOS switched-capacitor analog sampled-data direct-form recursive filters," *IEEE J. Solid-State Circuits,* vol. SC-14, pp. 1020–1033, Dec. 1979.

[2] R. Castello and P. R. Gray, "A high-performance micropower switched-capacitor filter," *IEEE J. Solid-State Circuits,* vol. SC-20, pp. 1122–1132, Dec. 1985.

[3] H.-S. Lee, D. A. Hodges, and P. R. Gray, "A self-calibrating 15 bit CMOS A/D converter," *IEEE J. Solid-State Circuits,* vol. SC-19, pp. 813–819, Dec. 1984.

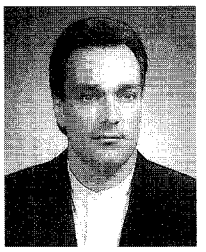
[4] P. W. Li, M. J. Chin, P. R. Gray, and R. Castello, "A ratio-independent algorithmic analog-to-digital conversion technique," *IEEE J. Solid-State Circuits,* vol. SC-19, pp. 828–836, Dec. 1984.

[5] T. L. Tewksbury, H.-S. Lee, and G. A. Miller, "The effects of oxide traps on the large-signal transient response of analog MOS circuits," *IEEE J. Solid-State Circuits,* vol. 24, pp. 542–544, Apr. 1989.

[6] T. L. Tewksbury and H.-S. Lee, "Characterization, minimization and modeling of transient threshold voltage shifts in MOSFETs," in *Proc. IEEE Custom Integrated Circuits Conf.,* San Diego, CA, May 1993, pp. 14.4.1–14.4.4.

[7] I. Lundström, S. Christensson, and C. Svensson, "Carrier trapping hysteresis in MOS transistors," *Phys. Status Solid (A),* vol. 1, no. 395, pp. 395–407, 1970.

[8] P. Rossel, H. Martinot, and A. Serhal, "Instabilities temporelles des transistors MOS-I," *Solid-State Electron.,* vol. 13, pp. 257–268, 1970.
[9] "Instabilities temporelles des transistors MOS-II," *Solid-State Electron.,* vol. 13, pp. 425–438, 1970.
[10] E. H. Snow and B. E. Deal, "Polarization effects in insulating films on silicon—A review," *Trans. Met. Soc. AIME,* vol. 242, pp. 512–523, Mar. 1968.
[11] P. Cuevas, "A simple explanation for the apparent relaxation effect associated with hot-carrier phenomenon in MOSFETs," *IEEE Electron Device Lett.,* vol. 9, pp. 627–629, Dec. 1988.
[12] D. Vuillaume, J. C. Bourgoin, and M. Lannoo, "Oxide traps in Si-SiO 2 structures characterized by tunnel emission with deep-level transient spectroscopy," *Phys. Rev. B,* vol. 34, pp. 1171–83, July 1986.
[13] D. J. DiMaria, "Trap creation in silicon dioxide produced by hot electrons," *J. Appl. Phys.,* vol. 65, no. 6, pp. 2342–2356, 1989.
[14] "Defects and impurities in thermal SiO_2," in *The Physics of SiO_2 and Its Interfaces,* S. T. Pantelides, Ed. New York: Pergamon, 1978, pp. 160–178.
[15] S. K. Lai and D. R. Young, "Effects of avalanche injection of electrons into silicon-dioxide—Generation of fast and slow interface states," *J. Appl. Phys.,* vol. 52, pp. 6231–6240, Oct. 1981.
[16] F.-C. Hsu and S. Tam, "Relationship between MOSFET degradation and hot-electron-induced interface-state generation," *IEEE Electron Device Lett.,* vol. EDL-8, pp. 50–52, Feb. 1984.
[17] M. Bourcerie, B. S. Doyle, J. C. Marchetaux, J. C. Soret, and A. Boudou, "Relaxable damage in hot-carrier stressing of n-MOS transistors—Oxide traps in the near interfacial region of the gate oxide," *IEEE Trans. Electron Devices,* vol. 37, pp. 708–717, Mar. 1990.
[18] B. S. Doyle, M. Bourcerie, J.-C. Marchetaux, and A. Boudou, "Relaxation effects in NMOS transistors after hot-carrier stressing," *IEEE Electron Device Lett.,* vol. EDL-8, pp. 234–236, May 1987.
[19] F. P. Heiman and G. Warfield, "The effects of oxide traps on the MOS capacitance," *IEEE Trans. Electron Devices,* vol. ED-12, pp. 167–178, Apr. 1965.
[20] S. Christensson, I. Lundström, and C. Svensson, "Low frequency noise in MOS transistors," *Solid-State Electron,* vol. 11, pp. 797–812, 1968.
[21] B. Balland and P. Pinard, "Interfacial tunnel interaction in MOS elements in depletion," *Phys. Status Solidi (A),* vol. 47, pp. 251–258, 1978.
[22] P. van Staa, H. Rombach, and R. Kassing, "Time-dependent response of interface states in indium phosphide metal-insulator-semiconductor capacitors investigated with constant-capacitance deep-level transient spectroscopy," *J. Appl. Phys.,* vol. 54, pp. 4014–4021, July 1983.
[23] V. Lakshmanna and A. S. Vengurlekar, "Logarithmic detrapping response for holes injected into SiO_2 and the influence of thermal activation and electric fields," *J. Appl. Phys.,* vol. 63, no. 9, pp. 4548–4554, 1988.
[24] T. L. Tewksbury, "Relaxation effects in MOS devices due to tunnel exchange with near-interface oxide traps," Ph.D. dissertation, Massachusetts Inst. Technol., 1992.
[25] H. Lakhardi, D. Vuillaume, and J. C. Bourgoin, "Spatial and energetic distribution of Si-SiO_2 near-interface states," *Phys. Rev. B,* vol. 38, pp. 13124–13132, Dec. 1988.
[26] I. Lundström and C. Svensson, "Tunneling to traps in insulators," *J. Appl. Phys.,* vol. 43, pp. 5045–5047, Dec. 1972.
[27] J. R. Chelikowsky and M. Schluter, "Electron states in α-quartz: A self-consistent pseudopotential calculation," *Phys. Rev. B,* vol. 15, no. 8, pp. 4020–4029, 1977.
[28] H. S. Fu and C. T. Sah, "Theory and experiments on surface $1/f$ noise," *IEEE Trans. Electron Devices,* vol. ED-19, pp. 273–285, Feb. 1972.
[29] W. Shockley and W. Read, "Statistics of the recombination of holes and electrons," *Phys. Rev.,* vol. 87, no. 835, pp. 835–842, 1952.
[30] M. Lax, "Cascade capture of electrons in solids," *Phys. Rev.,* vol. 119, pp. 1502–1523, Sept. 1960.
[31] D. V. Land and C. H. Henry, "Nonradiative recombination at deep levels in GaAs and GaP by lattice-relaxation multiphonon emission," *Phys. Rev. Lett.,* vol. 35, Dec. 1975.
[32] M. J. Kirton and M. J. Uren, "Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency $(1/f)$ noise," *Adv. Phys.,* vol. 38, no. 4, pp. 367–468, 1989.
[33] E. H. Nicollian and J. R. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology.* New York: Wiley, 1982.
[34] R. Jayaraman and C. G. Sodini, "A $1/f$ noise technique to extract the oxide trap density near the conduction band edge of silicon," *IEEE Trans. Electron Devices,* vol. 36, pp. 1773–1782, 1989.
[35] B. J. Gross and C. G. Sodini, "$1/f$ noise in MOSFETs with ultrathin gate dielectrics," in *IEDM Tech. Dig.,* Dec. 1992, pp. 881–884.
[36] Z. Celik-Butler and T. Y. Hsiang, "Spectral dependence of $1/f^{\gamma}$ noise on gate bias in n-MOSFETs," *Solid-State Electron.,* vol. 30, no. 4, pp. 419–423, 1987.
[37] N. Hwang, B. S. S. Or, and L. Forbes, "Tunneling and thermal emission of electrons from a distribution of deep traps in Si-O_2," *IEEE Trans. Electron Devices,* vol. 40, no. 6, pp. 1100, 1993.
[38] N. Shiono, M. Shimaya, and O. Nakajima, "Charge character of interface traps at the SiO_2 interface," *Appl. Phys. Lett.,* vol. 48, no. 17, pp. 1129–1131, 1986.
[39] A. Balasinski and T. Ma, "Enhanced electron trapping near channel edges in NMOS transistors," *IEEE Trans. Electron Devices,* vol. 39, pp. 1680–1686, July 1992.
[40] P. C. Dow, "An analysis of certain errors in electronic differential analyzers II—Capacitor dielectric absorption," *IRE Trans. Electron. Comput.,* pp. 17–22, Mar. 1958.
[41] J. W. Fattaruso, M. D. Wit, G. Warwar, K.-S. Tan, and R. K. Hester, "The effect of dielectric relaxation on charge-redistribution A/D converters," *IEEE J. Solid-State Circuits,* vol. 25, pp. 1550–1561, Dec. 1990.
[42] H. Koelmans and H. C. D. Graaff, "Drift phenomena in CdSe thin film FET's," *Solid-State Electron.,* vol. 10, pp. 997–1005, 1967.
[43] S. J. Prasad and S. J. T. Owen, "The InP-SiO_2 interface: Electron tunneling into oxide traps," in *Dielectric Films on Compound Semiconductors,* Electrochem. Soc., 1985, vol. 86-3, pp. 23–32.
[44] M. Declercq and P. Jespers, "Analysis of interface properties in MOS transistors by means of charge pumping measurements," *Rev. HF Acta. Techn. Belgica.,* vol. 9, 244–253, 1974.
[45] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions,* 10th ed. Washington, DC: Nat. Bureau Standards, 1972, ch. 5, p. 228.



Theodore L. Tewksbury, III (S'86–M'87) received the S.B. degree in architecture and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1983, 1987, and 1992, respectively.

In 1987 he joined Analog Devices, Inc., Wilmington, MA, as a Design Engineer for the Converter Group, where he worked on high-speed, high-resolution data acquisition circuits. Since 1992 he has been Senior Engineer in the Characterization Group, where he is involved in the modeling of advanced bipolar and submicron CMOS and BiCMOS devices. His research interests are in the areas of solid-state physics, device and circuit simulation, statistical modeling, and analog integrated circuit design.

Hae-Seung Lee (M'85–SM'92) was born in Seoul, South Korea, in 1955. He received the B.S. and M.S. degrees in electronic engineering from Seoul National University, Seoul, Korea, in 1978 and 1980 respectively, and the Ph.D. degree in electrical engineering from the University of California, Berkeley, 1984, where he developed self-calibration techniques for A/D converters.

In 1980 he was a member of the Technical Staff in the Department of Mechanical Engineering at the Korean Institute of Science and Technology, Seoul, where he was involved in the development of alternative energy sources. Since 1984 he has been with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, where he is now a Professor. Since 1985 he has acted as a Consultant to Analog Devices, Inc., Wilmington, MA, and M.I.T. Lincoln Laboratories, Lexington, MA. His research interests are in the areas of analog integrated circuits, early vision circuits, fabrication technologies, and solid-state sensors.

Dr. Lee is a recipient of the 1988 Presidential Young Investigator's Award. He has served on a number of technical program committees for various IEEE conferences, including the International Electron Devices Meeting, the International Solid-State Circuits Conference, the Custom Integrated Circuits Conference, and the IEEE Symposium on VLSI Circuits. Since 1992 he has been an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS.