A. V. Ferris-Prabhu

Theory of MNOS Memory Device Behavior

Abstract: This paper extends the direct-tunneling theory of MNOS memory device behavior to account for traps that are distributed both spatially and energetically. It shows that the Pulver and Dorda model is a special case of the Ross and Wallmark model, which is itself a restricted version of the more general theory proposed here.

A general equation for the total charge transfer is derived, for monoenergetic and for energetically uniform trap distributions, for high fields as well as low fields. The theory predicts the time dependence of the total charge transfer to be initially linear, then roughly logarithmic, and finally to reach saturation. There is no essential difference in the results whether the trap distribution is monoenergetic or energetically uniform. The operational dependence on characteristic parameters is investigated and found to be greatest for changes in the extent of spatial distribution of traps. The switching time varies inversely with trap density and tunnelling probability, and exponentially with the oxide thickness.

1. Introduction

There has been considerable interest in the use of metal-nitride-oxide-silicon (MNOS) devices as memory elements [1-10] and several theories have been advanced in order to describe the behavior of such devices. According to the direct tunnelling theory of Ross and Wallmark [1], charge transfer between silicon, and traps located in the silicon nitride and at the nitride-oxide interface, occurs by direct quantum mechanical tunnelling through the forbidden band of the oxide. Assuming a spatially uniform monoenergetic trap distribution, Ross and Wallmark predicted that the charge transfer varied as the logarithm of pulse width. Experimental evidence in support of those predictions has been presented by Ross and Wallmark, Svensson and Lundström [4] and Balk and Stephany [11]. Ross and Wallmark's results do not predict saturation nor are they valid for very short pulsewidths, since $\ln \tau < 0$ for $\tau < 1$.

An alternative approach by Dorda and Pulver [3] is also based on a direct tunnelling model. They assume an energetically uniform distribution located at a fixed distance from the oxide-nitride interface, and they predict that the charge transport should vary as a constant minus the negative exponential of the pulse width. They too have shown experimental evidence that appears to confirm their theory. Pulver and Dorda's model predicts physically satisfactory results but it is not particularly realistic to assume that all tunnelling occurs from a single plane at some fixed distance from the silicon-silicon dioxide interface.

Ross and Wallmark, in their expression for the charge transfer, let the nitride thickness l go to infinity since it is two orders of magnitude greater than other characteristic lengths appearing in their equation. As a consequence, the long time dependence is lost and their results are unable to predict saturation [12]. The correct short time behavior could have been recovered [12] had they not restricted their result to times t greater than some characteristic value t_0. For $t > t_0$ they correctly neglected the exponential integral term appearing in their result. However, they then defined a switching time t' by an expression valid only for $t' < t_0$ leading to an inconsistency in their definition which obscures its meaning [13].

The model of Dorda and Pulver leads to an expression for charge transfer that is essentially the integrand of Ross and Wallmark's expression. This was shown [13] to be so because of Dorda and Pulver's assumption that for effective purposes all the traps could be considered to lie at some fixed distance from the oxide-nitride interface. For very short pulse widths both the corrected Ross and Wallmark expression and Dorda and Pulver's results show the same initial linear time dependence [13].

Lundström and Svensson [4], using a general model for charge accumulation in a double insulator structure, have again obtained a logarithmic time dependence. Their result is a direct and necessary consequence of their assumption that the current is exponentially depen-

dent on the electric field. Moreover, experimentally one can get reasonable fits to log I vs V or I/V or $\sqrt{V}$ over many orders of magnitude in the current and over only one order of magnitude in the voltage. Thus log I vs V will fit many experimental results.

In the present paper the direct-tunnelling theory is generalized and applied to a trap distribution that is uniform, both spatially and energetically. Section 2 develops the model and Section 3 formulates the charge transport equation. This is solved in Section 4 for a monoenergetic trap distribution (for comparative purposes) and in Section 5 for an energetically uniform trap distribution. Section 6 investigates the operational dependence on characteristic parameters, Section 7 presents some experimental results, and conclusions are given in Section 8.

2. Model

The phenomenon of charge storage at the interface between two dielectrics enables layered-insulator FETs to function as memory elements. Depending on the polarity of the stored charge, the turn-on voltage of the device acquires a high or a low value, providing the two states needed for binary operation. For charge storage to occur it is necessary that the divergence of the current across the interface be nonzero. This requires the existence of traps at and near the interface and also that the current through one insulator be unequal to current through the other, at least initially. The existence of traps has been verified by Kendall [14] and others [15], although their energetic and spatial distribution has not been resolved unequivocally. The nonequality of insulator currents is achieved by using a thin oxide of about 20 to 30 Å and a thicker nitride of from 200 to 500 Å. Charge transfer through oxides of 35 Å or less occurs by direct quantum mechanical tunnelling through the oxide conduction band, whereas for thicker oxides charge transfer is effected by Fowler-Nordheim tunnelling [1,4]. Charge transfer through the nitride can occur [16,17] by field enhanced thermal excitation (Poole-Frenkel), by field ionization of trapped electrons into the conduction band of the nitride, and by the hopping of thermally activated electrons between isolated traps. The first two of these processes require electric fields of 10^7 V/cm or greater [16,17]. For typical operating voltages of 20 V, the field across a 500-Å nitride is 4×10^6 V/cm. Thus only the trap hopping mechanism is likely to occur. However, at least initially, tunnelling through the oxide is a much faster process, permitting the nitride conduction to be neglected [1-4]. Near saturation, when steady state conditions are being reached, the nitride current will of course assume a larger relative role.

The present model considers the case for which the oxide current is dominant and due to direct tunnelling between the silicon and traps lying at or near the oxide-nitride interface.

The charge transfer during an interval of time dt, between the silicon and traps with energies in an interval E, $E + dE$ lying in the nitride at a distance between x and $x + dx$ from the silicon-silicon dioxide interface, is

$$dQ(x,E,t) = q\phi(x,E,t)\, dx\, dE\, dt\,, \tag{1}$$

where q is the unit of charge and ϕ is the particle flux. The particle flux is given by

$$\phi(x,E,t) = \omega(x,E,t)\mathcal{N}(x,E,t)\,, \tag{2}$$

where ω is the tunnelling probability of the particles; at time t there are $\mathcal{N}$ particles having energy E at location x. The number of particles decreases according to the usual decay law,

$$d\mathcal{N}(x,E,t)/dt = -\omega(x,E,t)\ \mathcal{N}(x,E,t)\,. \tag{3}$$

The tunnelling probability is a rather complicated quantity that depends on the position of the particle, its energy (which is a function of its position and the gate potential), and the height of the potential barrier (which depends on the thickness of the insulating layers, the gate potential and the amount of charge at the oxide-nitride interface). Since the amount of interfacial charge changes with time, neglecting its effect is equivalent to neglecting the time dependence of the tunnelling probability. It seems reasonable to use the tunnelling probability appropriate to a rectangular barrier as a first approximation*, i.e.,

$$\omega(x,E,t) = \omega_0 \exp\,(-x/\lambda)\,, \tag{4}$$

where ω_0, which has the units of reciprocal time, and λ, which has the units of length, are adjustable parameters related to the details of the potential barrier and the particle energy.

With these simplifications, Eq. (4) may be solved, yielding

$$\mathcal{N}(x,E,t) = \mathcal{N}(x,E,0) \exp\,[-\omega(x)t]\,. \tag{5}$$

The number of particles, $\mathcal{N}$, at time zero is related to the number of traps, N, by the relation

$$\mathcal{N}(x,E,0) = N(x,E)\theta\,, \tag{6}$$

where θ is the probability (one or zero) that the trap contains a particle. It is assumed for simplicity that θ is unity. Then

$$\mathcal{N}(x,E,0) = N(x,E)\,. \tag{7}$$

The form of the trap distribution is not known but it seems reasonable that the number of traps at any level E and location x can be written as

*A significant limitation of this assumption is that for fields above a certain value the charge transport equation no longer contains a field dependent term.

$$N(x,E) = N_0 f(E) e^{-\alpha(E)x}, \tag{8}$$

where $\alpha(E)$ is unknown. Since the tunnelling probability drops off very rapidly with distance, we shall approximate $e^{-\alpha(E)x}$ by unity for $x < x_m$ where x_m is an adjustable parameter [13], and by zero for $x > x_m$. This is, of course, the same as assuming a uniform spatial distribution for $x < x_m$. Since the extent of spatial distribution of traps x_m is identified as the maximum distance from which tunnelling can occur, and since x_m is not much greater than about 30 to 35 Å, it is unlikely that any significant error should result by replacing the actual spatial distribution by a uniform distribution.

The physical significance of x_m is this: If traps are located with uniform density in the entire nitride layer, a very small gate potential will bring the more distant traps up to the silicon conduction band edge and charge transfer will occur. However, no shift in flat band voltage is measureable when the gate potential is less than 5 V [12]. This suggests that beyond some distance x_m from the semiconductor-insulator interface, either the trap density or the charge transfer is negligible.

However, if saturation is caused by charge build-up, it implies that sufficient charge has been stored in the traps to negate the effect of the applied voltage. If the traps are fairly close to each other, it is not possible for the ones nearest the interface to be filled without at least partially filling the ones fairly near to themselves. Therefore, traps extending some distance away from the interface will also be filled to some degree before the charge build-up is sufficient to prevent further charge transfer. It is physically reasonable to interpret x_m as this distance. In effect, traps beyond some distance x_m do not play a substantial role in the charge transfer between silicon and insulator.

It seems reasonable to assume a random distribution of the dislocations and other defects at the oxide-nitride interface that lead to the existence of traps. Thus the trap distribution $f(E)$ may well be gaussian as suggested in Fig. 1. However, the exact distribution is not known unequivocally and is being investigated in this laboratory and elsewhere [8,14].

Finally, from Eqs. (2), (4), (5), (7) and (8):

$$\mathcal{N}(x,E,t) = N_0 f(E) e^{-\omega(x)t}, \tag{9}$$

and

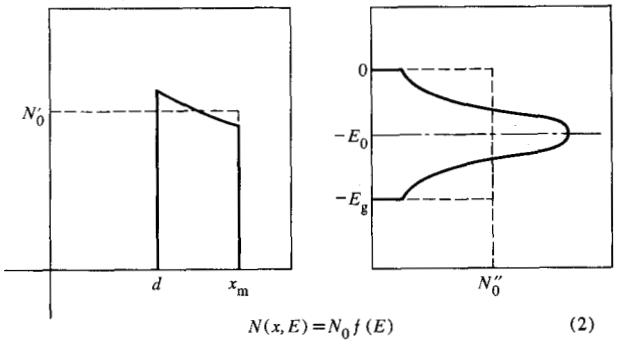
$$\phi(x,E,t) = N_0 f(E) \omega(x) e^{-\omega(x)t}. \tag{10}$$

As shown in Fig. 2, where only a single trap level is indicated for clarity, the trap level has to be raised to a level $|q\ \psi_s|$ above the silicon conduction band edge for tunnelling to be possible. This minimum distance from the silicon-silicon dioxide interface is x_0.

$$N(x,E) = N_0 \exp[-\alpha(E)x] \exp -[(E+E_0)/kT]^{2'} \quad (1)$$

$$d \le x \le x_m$$

$$-E_g \le E \le 0$$

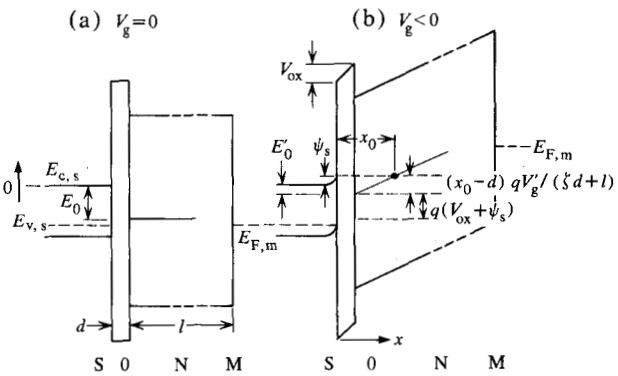


$$N(x,E) = N_0 f(E) \tag{2}$$

$$f(E) = \begin{cases} \delta(E+E_0) \\ 1 \end{cases} \tag{3}$$

$$\phi(x,E\ t) = N_0 f(E) \omega(x) \exp[-\omega(x)t] \tag{4}$$

Figure 1 Trap distribution.



$$E'_0 = E_0 + q\,V_{ox} + q\,\psi_s$$

$$V_{ox} = [\zeta d/(\zeta d+l)]\ V'_g$$

$$x_0 = \begin{cases} (1-\zeta)d + (\zeta d+l)\,[E_0/(qV'_g)], & qV'_g \le [(\zeta d+l)/\zeta d]\,E_0 \\ d, & qV'_g > [(\zeta d+l)/\zeta d]\,E_0 \end{cases}$$

$$V'_g = V_g - \phi_{ms} - \psi_s$$

Figure 2 Simplified energy band diagram for MNOS memory devices.

By invoking the requirement of electric flux continuity across the oxide-nitride interface, neglecting the charge trapped at this interface, and summing all potential drops across the structure, a simple geometrical calculation shows that

$$x_0 = (1-\zeta)d + (\zeta d + l)|E|/q\,V_g', \tag{11}$$

where $\zeta = \epsilon_n/\epsilon_{ox}$, ϵ_n and ϵ_{ox} being the permittivities of the nitride and oxide of thickness l and d respectively, and $|E|$ is the magnitude of the difference between the trap level and the silicon conduction band edge. Note that we

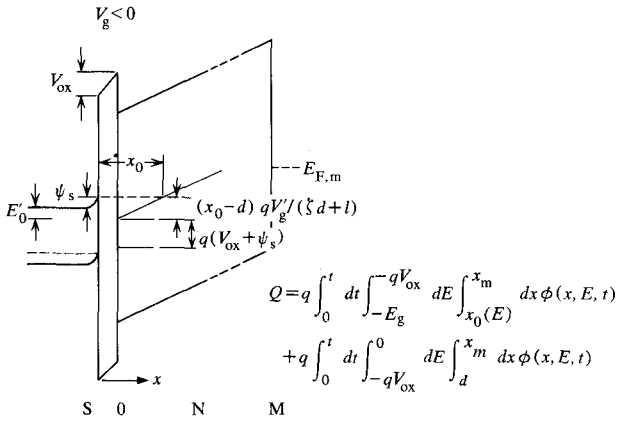


Figure 3 Charge transport equation and illustration.

have adopted the convention that energies above the conduction band edge are considered positive and energies below it are taken to be negative.* And

$$V_g' = V_g - \phi_{ms} - \psi_s\,, \tag{12}$$

where ϕ_{ms} is the metal-semiconductor work function difference, ψ_s is the silicon surface potential, and V_g is the signed gate potential. For electrons, q is negative, making $q\,V_g$ positive for a negative gate voltage.

The minimum tunnelling distance x_0 cannot be less than the separation d between the oxide-silicon interface and the traps at the oxide-nitride interface, nor is it consistent to let x_0 exceed x_m, the maximum distance up to which traps effectively exist. The first restriction provides a definition of high field and the second defines a threshold gate voltage below which tunnelling cannot occur. We therefore define the high-field gate voltage by

$$qV_{gH}' \geq (1 + l/\zeta d)E_0 \tag{13}$$

and the threshold gate voltage by

$$qV_{gL}' \geq \frac{\zeta d}{x_m + (\zeta - 1)d}\,(1 + l/\zeta d)E_0\,, \tag{14}$$

where E_0 is the magnitude of the energy level of the deepest trap, measured from the silicon conduction band edge, available for electron tunnelling.
Thus

$$x_0(E,qV_g') = \begin{cases} (1-\zeta)d + (\zeta d + l)\,\dfrac{|E|}{qV_g'}, & \\ & V_{gL}' \leq V_g' \leq V_{gH}' \\ d, & V_g' \geq V_{gH}'\,. \end{cases} \tag{15}$$

*By considering energies below the silicon valence band edge to be positive and energies above it to be negative, the present analysis can be equally well applied to hole tunnelling with appropriate changes in the effective mass, etc.

3. Charge transport equation

From Eqs. (1) and (10),

$$dQ = qN_0 f(E)\,\omega(x)e^{-\omega(x)t}dx\,dE\,dt\,. \tag{16}$$

The trap levels can be divided into two groups: those which lie up to a depth $q\,V_{ox}$ below the silicon conduction band edge, and those which lie deeper. Upon application of a gate potential sufficient to create a voltage drop across the oxide of value V_{ox}, the first group gets raised so that all states at the oxide-nitride interface can "communicate" with the silicon. For the second group, the nearest states that are raised sufficiently to communicate with the silicon lie at distance $x_0(E)$. Thus the total charge transport can be expressed by (see Fig. 3)

$$\frac{Q(t)}{qN_0} = \int_{-E_g}^{-qV_{ox}} dE\,f(E)\int_{x_0(E)}^{x_m} dx\,\omega(x)\int_0^t dt'\,e^{-\omega(x)t'} + \int_{-qV_{ox}}^{0} dE\,f(E)\int_d^{x_m} dx\,\omega(x)\int_0^t dt'\,e^{-\omega(x)t'}\,. \tag{17}$$

It is more convenient to write this in an alternative form as

$$\frac{Q(t)}{qN_0} = \int_{-E_g}^{0} dE\,f(E)\int_d^{x_m} dx\,\omega(x)\int_0^t dt'\,e^{-\omega(x)t'} - \int_{-E_g}^{-qV_{ox}} dE\,f(E)\int_d^{x_0(E)} dx\,\omega(x)\int_0^t dt'\,e^{-\omega(x)t'}\,. \tag{18}$$

The first term represents contributions from all trap levels, and the second term subtracts contributions from the lower trap levels located between the oxide-nitride interface at d and the minimum tunnelling distance $x_0(E)$.

Equation (17) or (18) can be written more generally as

$$Q(t) = qN_0\int_E dE\,f(E)\int_x dx[1 - e^{-\omega(x)t}]\,. \tag{19}$$

Ross and Wallmark's assumption of a single trap level E_0eV below the silicon conduction band edge is

$$f(E) = \delta(E + E_0) \tag{20}$$

which, when inserted in Eq. (19), gives the Ross and Wallmark charge transport equation

$$Q_{RW}(t) = Q(t)\,\delta(E + E_0) = qN_0\int dx[1 - e^{-\omega(x)t}]\,. \tag{21}$$

With the further assumption that all tunnelling originates from a single plane at x_p, Pulver and Dorda's charge transport equation is obtained. Thus:

$$Q_{PD}(t) = Q_{RW}(t)\delta(x - x_p) = qN_0[1 - e^{-\omega(x)t}]\,. \tag{22}$$

It is evident, as shown in more detail elsewhere [13] that the charge transport expressions of Pulver and Dorda, as those of Ross and Wallmark are contained in the more general expression of Eqs. (17), (18) or (19).

The temporal and spatial integrals of Eq. (18) are readily evaluated, giving

$$\frac{Q(\tau)}{qN_0\lambda} = \int_{-E_g}^{0} dE f(E)\{K_1 + E_1(\tau) - E_1(e^{-K_1}\tau)\} - \int_{-E_g}^{-qV_{0x}} dE f(E)\{K_2 + E_1(\tau) - E_1(e^{-K_2}\tau)\}, \tag{23}$$

where

$$\tau = \omega_0 t\, e^{-d/\lambda}, \tag{24}$$

$$K_1 = (x_m - d)/\lambda, \tag{25a}$$

and

$$K_2 = K_2(E) = (x_0(E) - d)/\lambda. \tag{25b}$$

From Eqs. (13), (14) and (15), it is evident that

$$0 \le K_2(E) \le K_1. \tag{26}$$

For gate voltages near the threshold, $K_2 \approx K_1$, and for voltages near the high field value, $K_2 \approx 0$.

The function $E_1(\tau)$ is related to the exponential integral and may be approximated as [18]

$$E_1(\tau) = \begin{cases} -\ln\tau + \sum_{n=0}^{5} a_n \tau^n, & 0 \le \tau \le 1 \\ g(\tau), & \tau \ge 1 \\ 0, & \tau \gg 1, \end{cases} \tag{27}$$

where

$a_0 = -0.577$,

$a_1 = 1$,

$a_2 = -0.25$,

$a_3 = 0.055$,

$a_4 = -0.009$,

$a_5 = 0.001$,

and

$$g(\tau) = \frac{e^{-\tau}}{\tau} \cdot \frac{\tau^2 + 2.33\tau + 0.25}{\tau^2 + 3.33\tau + 1.68} \tag{28}$$

decreases to zero very rapidly as τ increases.

Equation (23) contains the complete time dependence for any form of trap energy distribution. It is convenient to write this expression again as

$$Q(\tau)/q\,N_0\lambda = Q_{high} - Q_{corr}, \tag{29}$$

where the high-field contribution is

$$Q_{high} = \int_{-E_g}^{0} dE f(E)\{K_1 + E_1(\tau) - E_1(e^{-K_1}\tau)\} \tag{30}$$

and the low field correction is

$$Q_{corr} = \int_{-E_g}^{-qV_{ox}} dE f(E)\{K_2 + E_1(\tau) - E_1(e^{-K_2}\tau)\}. \tag{31}$$

An alternative way of writing Eqs. (30) and (31) is

$$Q_{high} = \int_{-E_g}^{0} dE f(E)\{[(x_m - d)/\lambda] + E_1(t/t_d) - E_1(t/t_m)\}, \tag{30a}$$

and

$$Q_{corr} = \int_{-E_g}^{-qV_{0x}} dE f(E)\langle\{[x_0(E) - d]/\lambda\} + E_1(t/t_d) - E_1(t/t_0)\rangle, \tag{31b}$$

where

$$1/t_d = \omega_0\, e^{-d/\lambda}, \tag{32a}$$

$$1/t_m = \omega_0\, e^{-x_m/\lambda}, \tag{32b}$$

and

$$1/t_0 = \omega_0\, e^{-x_0/\lambda}. \tag{32c}$$

4. Monoenergetic trap distribution

For a single trap level lying E_0 eV below the silicon conduction band edge,

$$f(E) = \delta(E + E_0) \tag{33}$$

and Eq. (23) becomes

$$Q(\tau) = [K_1 + E_1(\tau) - E_1(e^{-K_1}\tau)] - [K_2(-E_0, V_g') + E_1(\tau) - E_1(e^{-K_2}\tau)]. \tag{34}$$

For $V_g' \ge V'_{gH}$, the second set of terms vanishes giving

$$\frac{Q(\tau)}{qN_0\lambda} = \begin{cases} \sum_{1}^{5} (1 - e^{-nK_1})\, a_n \tau^n, & 0 \le \tau \le 1 \\ \ln\tau + g(\tau) + \gamma - \sum_{1}^{5} e^{-nK_1}\, a_n \tau^n, & 1 \le \tau \le e^{K_1} \\ K_1, & \tau > e^{K_1}. \end{cases} \tag{35}$$

Figure 4 shows Eq. (34) for $V_g' \ge V'_{gH}$.

For $V_g' < V'_{gH}$, the second set of terms in Eq. (34) must be included. Thus

$$Q_{corr} = \begin{cases} \sum_{1}^{5} [1 - e^{-nK_2(-E_0, V_g')}]\, a_n \tau^n, & 0 \le \tau \le 1 \\ \ln\tau + g(\tau) + \gamma - \sum_{1}^{5} e^{-nK_2}\, a_n \tau^n, & 1 \le \tau \le e^{K_2} \\ K_2(-E_0, V_g'), & \tau > e^{K_2}. \end{cases} \tag{36}$$

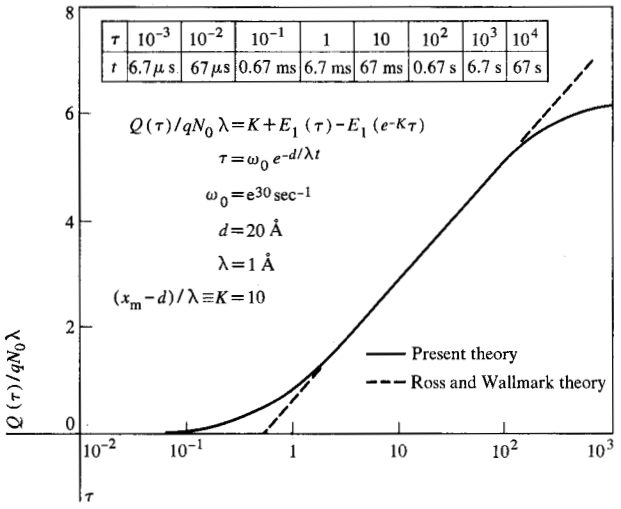


Figure 4 Charge transport vs time.

Subtracting (36) from (35) gives, for $V_g' < V'_{gH}$

$$\frac{Q(\tau)}{qN_0\lambda} = \begin{cases} \sum_1^5 (e^{-nK_2} - e^{-nK_1})\, a_n \tau^n, & 0 < \tau \le e^{K_2} \\ -K_2(-E_0, V_g') + \ln\tau + g(e^{-K_2}\tau) + \gamma & \\ -\sum e^{-nK_1}\, a_n \tau^n, & e^{K_2} \le \tau \le e^{K_1} \\ K_1 - K_2(-E_0, V_g'), & \tau > e^{K_1}. \end{cases} \quad (37)$$

Recalling the definitions of K_2 and V'_{gH}, we see that for V_g' slightly less than V'_{gH}, $K_2 \approx 0$. With this value of the gate potential, the minimum tunnelling distance is very slightly greater than the oxide thickness, i.e., traps located very near the oxide-nitride interface can communicate with the silicon conduction band, and the charge transport should be almost that corresponding to V'_{gH}. Indeed, for $V_g' = V'_{gH}$ Eq. (39) reduces to Eq. (37).

From the definition of V'_{gL}, it is seen that for V_g' slightly greater than V'_{gL}, $K_2 \approx K_1$. The minimum tunnelling distance x_0 is then only slightly less than the maximum tunnelling distance x_m and therefore not many traps can contribute to charge transport. Hence Q will be small. Equation (32) bears this out when $K_2 \approx K_1$ is inserted.

Examination of Eq. (37) shows that in the interval $0 \le \tau \le 1$, Q is initially linear, then gradually decreases from linearity. In the interval $[e^{K_2}, e^{K_1}]$, $g(\tau)$ is initially dominant but very rapidly decreases to zero. The logarithmic term then dominates until for larger values of τ the summation term exerts a stronger effect, pulling Q towards a constant value which is reached for $\tau > e^{K_1}$. The theory predicts that saturation sets in at $\tau \ge e^{K_1}$. Since $K_1 = (x_m - d)/\lambda$, x_m governs onset of saturation. We recall that x_m had been defined earlier as the maximum distance from the semiconductor-oxide interface to where the trap density is significant. Traps located beyond x_m are assumed not to contribute to charge transfer. Experiments indicate that saturation ensues after about 60 seconds [13]. Using the definition of τ, with $\omega_0 = e^{30}$/sec [3], and $\lambda = 1$ [1], we find that

$$x_m = \lambda \ln(\omega_0 t_{sat}) = 34 \text{ Å}. \quad (38)$$

It is evidently not physically justifiable to let $x_m = l$ and still less so to let $x_m \to \infty$. Since tunnelling probability decreases exponentially with distance, it may seem that it is unnecessary to cut off the trap distribution at x_m. But if we replace x_m by l, Eq. (35) or (37) shows that saturation ensues only after $\tau \ge e^{l/\lambda}$. This is evidently an unrealistic value, and becomes more so if l is replaced by infinity. It is because Ross and Wallmark assumed that l could be replaced by infinity that their results cound not predict a saturation time, nor could they predict the switching time in a self-consistent way [13].

5. Energetically uniform trap distribution

For traps distributed uniformly opposite the forbidden band of silicon,

$$f(E) = \begin{cases} 1, & -E_g \le E \le 0 \\ 0, & \text{other } E. \end{cases} \quad (39)$$

Then for $V_g' \ge V'_{gH}$,

$$\frac{Q(\tau)}{qN_0\lambda} = \int_{-E_g}^{0} dE[K_1 + E_1(\tau) - E_1(e^{-K_1}\tau)]$$

$$= E_g[K_1 + E_1(\tau) - E_1(e^{-K_1}\tau)]. \quad (40)$$

For low fields, it is necessary to evaluate

$$Q_{corr} = \int_{-E_g}^{-qV_{ox}} dE[K_2(E, V_g') + E_1(\tau) - E_1(e^{-K_2}\tau)]. \quad (41)$$

A straightforward evaluation shows that

$$\int_{-E_g}^{-qV_{ox}} dE\, K_2(E, V_g') = \frac{1}{2}\frac{\lambda}{\zeta d + l} q V_g' K_2^2(-E_g, V_g') \quad (42)$$

and

$$\int_{-E_g}^{-qV_{ox}} dE\, E_1(\tau) = \frac{\lambda}{\zeta d + l} q V_g' K_2(-E_g, V_g)\, E_1(\tau). \quad (43)$$

The third term in Eq. (41) turns out to be a form of Van de Hulst's integral [19],

$$E_1^2(\xi) \equiv \int_0^{\xi} d\eta \frac{E_1(\eta)}{\eta}. \quad (44)$$

Specifically,

$$\int_{-E_g}^{-qV_{ox}} dE\, E_1(e^{-K_2}\tau) = \frac{\lambda}{\zeta d + l} q V_g'[E_1^2(e^{-K_2}\tau) - E_1^2(\tau)]. \quad (45)$$

Inserting Eqs. (42), (43) and (45) into Eq. (41) gives

$$Q_{\text{corr}} = \frac{\lambda}{\zeta d + l} q V_g' \left[\frac{1}{2} K_2^2 + E_1(\tau) + E_1^2(\tau) - E_1^2(e^{-K_2}\tau) \right]. \quad (46)$$

Hence for $V_g' < V_{gH}'$,

$$Q(\tau)/qN_0\lambda = \left[K_1 - \frac{1}{2} \frac{\lambda}{\zeta d + l} \frac{qV_g'}{E_g} K_2^2(-E_g, V_g) \right] + \left\{ \left[1 - \frac{\lambda}{\zeta d + l} \frac{qV_g'}{E_g} K_2(-E_g, V_g') \right] E_1(\tau) - E_1(e^{-K_1}\tau) \right\} - \frac{\lambda}{\zeta d + l} \frac{qV_g'}{E_g} [E_1^2(\tau) - E_1^2(e^{-K_2}\tau)]. \quad (47)$$

It is clear that for $V_g' = V_{gH}' = \frac{\zeta d + l}{\zeta d} E_g$, $K_2 = 0$, and Eq. (47) reduces to Eq. (40). This is similar to the monoenergetic case.

Since it is difficult to interpret Eq. (47) readily, another approximate form is obtainable by first using the series representations for $E_1(\tau)$ and then performing the integrations indicated in Eq. (41). Doing so yields the result, for $V_g' < V_{gH}'$,

$$\frac{Q(\tau)}{qN_0\lambda E_g} = \begin{cases} \sum_{n=1}^{5} \left[(1 - e^{-nK_1}) - C\left(1 - \frac{1 - e^{-nK_2}}{nK_2}\right) \right] a_n \tau^n, & 0 \le \tau \le 1 \\ (1 - C)[g(\tau) + \ln \tau + \gamma] + \sum_{n=1}^{5} \left[C \frac{1 - e^{-nK_2}}{nK_2} - e^{-nK_1} \right] a_n \tau^n, & 1 < \tau \le e^{K_2} \\ (1 - C)[g(\tau) + \ln \tau + \gamma] - \sum_{n=1}^{5} e^{-nK_1} a_n \tau^n - \frac{1}{2} CK_2 + Cg(e^{-K_2}\tau), & e^{K_2} < \tau \le e^{K_1} \\ (1 - C)\, g(\tau) + K_1 - g(e^{-K_1}\tau) - \frac{1}{2} CK_2 + Cg(e^{-K_2}\tau), & \tau < e^{K_1}, \end{cases} \quad (48)$$

where

$$C = \frac{\lambda}{\zeta d + l} \frac{qV_g'}{E_g} K_2,$$

and

$$K_2 \equiv K_2(-E_g, V_g') = \frac{\zeta d}{\lambda} \left[1 - \frac{\zeta d + l}{\lambda} \frac{E_g}{qV_g'} \right].$$

When the lowest trap level is raised to the silicon conduction band edge, i.e., for $V_g' \ge V_{gH}'$, $K_2 = 0$ and there is no longer any voltage dependence, and Eq. (48) reduces to

$$\frac{Q(\tau)}{qN_0\lambda E_g} = \begin{cases} \sum_{1}^{5} (1 - e^{-nK_1})\, a_n \tau^n, & 0 \le \tau \le 1 \\ \ln \tau + g(\tau) + \gamma - \sum e^{-nK_1} a_n \tau^n, & 1 < \tau \le e^{K_1} \\ K_1, & \tau > e^{K_1}. \end{cases} \quad (49)$$

This result is identical to the monoenergetic high field case.

These results suggest that for high fields no essential difference is introduced by assuming either a monoenergetic distribution or an energetically uniform one. For low fields the latter model predicts a slightly more complicated dependence on the gate voltage than the monoenergetic model.

The high field regime as defined in Section 2, Eq. (13) is shown in the next section to occur for gate voltages greater than 15 V. For a 500 Å nitride this corresponds to an electric field of 2.8×10^6 V/cm. At a typical 20 V gate voltage, the electric field strength is 4×10^6 V/cm, at which level the very low conduction that occurs is due to the hopping of thermally activated electrons from one isolated trap to another [16,17]. Thus it is not expected, for short pulses, that conduction in the nitride will play a significant role, compared to the oxide current.

6. Operational dependence on characteristic parameters

To investigate the operational dependence on characteristic parameters such as insulator thicknesses [20,21] we next consider the shift in flatband voltage

$$V_s(t) = (l/\epsilon_n)\, Q\ (t). \quad (50)$$

From Eq. (13), with $\zeta = 1.7$, $l \approx 500$ Å and $d \approx 25$ Å, it is found that $V'_{gH} \approx 15$ V. This value is less than typical operating voltages of 20 V or more. Thus it is appropriate to consider only the high field expression for $Q(t)$. In that case, using the notations of Eqs. (30a), (31a) and (32), we have

$$V_s(t) = (l/\epsilon_n) qN_0\lambda\{[x_m - d)/\lambda] + E_1(t/t_0) - E_1(t/t_m)\}, \quad (51)$$

where x_0 has been replaced by d.

It is evident that V_s increases with N_0, l and x_m. The dependence on the other parameters can be inferred from the approximate relation

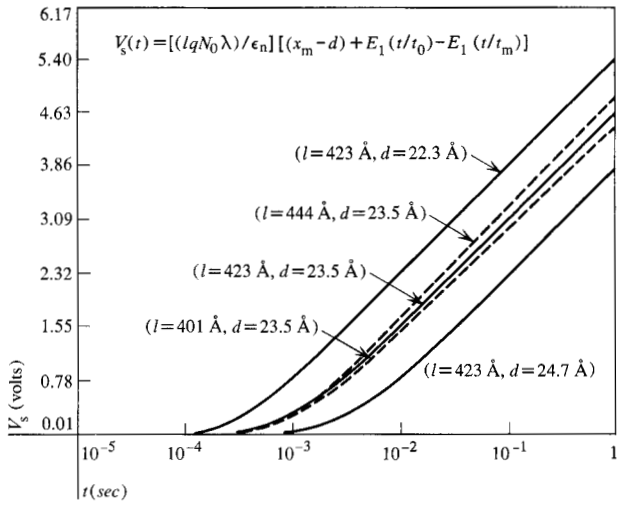


Figure 5 Shift in flatband vs pulse width for different applied voltages ($q = 1.6 \times 10^{-19}$ coul, $N_0 = 6 \times 10^{19}$ cm^{-3}, $\epsilon_n = 6.2 \times 10^{-13}$ F/cm, $\omega_0 = 10^{13}$/sec, $\lambda = 1$ Å, $x_m = 35$ Å, $t_0 = e^{d/\lambda}/\omega_0$, $t_m = e^{x_m/\lambda}/\omega_0$).

$$V_s(t) \approx \frac{qN_0\lambda l}{\epsilon_n} \begin{cases} \omega_0 t e^{-d/\lambda}, & t < t_o \quad (52a) \\ \ln(\omega_0 t) - d/\lambda, & t_0 < t < t_m \quad (52b) \\ (x_m - d), & t_m < t, \quad (52c) \end{cases}$$

where terms containing $e^{-x_m/\lambda}$ have been neglected because they are very much smaller than $e^{-d/\lambda}$.

Equation (52) shows that V_s increases as the tunnelling probability ω_0 and the de Broglie wavelength λ increase. But V_s decreases as the oxide thickness d increases. This is so because, as the oxide gets thicker, less charge is transferred and so the shift in flatband voltage is consequently decreased.

To verify these inferences Eq. (51) was evaluated for nominal and excursionary values of the parameters l, d, ω_0, λ and x_m. Typical results are shown in Fig. 5. It was found that variations in λ have the greatest effect, followed by oxide thickness d and tunnelling probability ω_0. Changes in x_m and l have very little effect, as expected.

Now, the de Broglie wavelength λ as well as the tunnelling probability ω_0 depend quite sensitively on the height and shape of the potential barrier. The latter is affected in turn by the applied voltage, the trap levels and the oxide thickness. It would thus appear that variations in oxide thickness may exert a dominant effect on device operational characteristics. This has also been suggested by Ross [20], Chou [21], and Goodman [22].

Consider next the dependence of the switching time t' on trap density, tunnelling probability, etc. This may be found by inverting Eq. (51). It is sufficient, however, merely to invert Eq. (52a) giving

$$t' = \left(\frac{\epsilon_n e^{d/\lambda}}{qN_0\lambda l\omega_0}\right) V_s', \quad (53)$$

where V_s' is some prescribed value for the shift in flatband voltage at which, by definition, switching occurs. After rewriting Eq. (53) as

$$t' = \left(\frac{\epsilon_n V_s'}{q}\right) \frac{e^{d/\lambda}}{N_0\lambda l\omega_0}, \quad (54)$$

it is evident that t' decreases as the trap density, de Broglie wavelength, nitride thickness and tunnelling probability increase, and as the oxide thickness decreases. However, the dependences on d and λ are quite strong.

At operational voltages all trap levels are raised well above the silicon conduction band edge and the present theory assumes that the tunnelling probability does not change. This is, of course, a simplification. The pulse amplitude dependence given by Ross and Wallmark [1] is correct only for the case where the gate voltage is insufficient to raise the trap level above the silicon conduction band edge. Since such voltages are much less than operational voltages, the exponential dependence on pulse amplitude is then of limited validity.

A typical value for t' can be calculated from Eq. (53). Assume $N_0 = 10^{20}$ cm^{-3} [1], $\lambda = 1$ Å [2], and $\omega_0 = 10^{13}$ sec^{-1} [3]. Then with $\epsilon_n = 6.2 \times 10^{-13}$ F-cm^{-1}, $q = 1.6 \times 10^{-19}$ coul, $d = 20$ Å and $l = 500$ Å, if 0.5 V is the value of shift in flatband voltage, which can be readily determined, we find that $t' = 18$ μs. The exact value cannot be calculated without obtaining exact values for N_0, λ and ω_0. It is clear that the switching time t' has meaning only in the context of the sensitivity of the apparatus used to measure V_s'. Recent measurements [23] have shown that the charge transferred during application of the gate potential, decays to about one-fifth of its initial value within microseconds after the cessation of the pulse. Hence it might be more correct from an operational point of view to define switching time as the minimum pulse, for a given applied potential, that will cause the shift in flatband voltage to remain at some prescribed value for several seconds after cessation of the pulse.

7. Experimental results

The previously derived equations can be used to calculate the values of various characteristic parameters. Recalling the definition of V_{gL} as the threshold or lowest gate voltage for which a shift in flatband voltage is measurable, the extent of spatial distribution of traps, from Eq. (13), is given by

$$x_m = (1 - \zeta)\, d + (\zeta d + l)\, (E_0/qV_{gL}). \quad (56)$$

Also, from Eqs. (52a) and (52c), the trap density and the tunnelling probability can be expressed as

$$N_0 = V_{s,\,sat}/[(l/\epsilon_n)\ q\ \lambda\ (x_m - x_0)] \tag{57}$$

and

$$\omega_0 = \frac{\partial V_s/\partial t}{(l/\epsilon_n)\ qN_0\ \lambda \exp\ (-x_0/\lambda)}. \tag{58}$$

Equations (56) through (58) provide means of verifying the characteristic parameters ω_0, N_0, and x_m, directly from experimental data. The parameters λ and E_0 are not amenable to extraction from the present formulation and need to be obtained from other experiments.

We used 2 Ω-cm p-silicon with a 20 Å thick layer of dry thermal oxide. A nitride layer 200 Å thick was grown at 900°C using a 1:10 silane-to-ammonia ratio. The gate material was rf-heated aluminum.

All measurements were taken at 30°C, and each time after a pulse was applied, the flatband was brought back to the original condition. Measurements were carried out on several devices and representative values are shown in Fig. 6.

It was found that the lowest applied voltage for which a shift in flatband voltage was measurable was 5 V. Using Eq. (7), x_m was calculated to be 32.8 Å, which is in substantive agreement with the well known fact that tunnelling is negligible for distances greater than about 35 Å.

At 25 V gate amplitude, for a succession of equal pulses, the saturation value V_{sat} of the flatband voltage is 6.2 V. With $\zeta = 1.7$, $(\psi_s + \phi_{ms}) = 1$ V, $q = 1.6 \times 10^{-19}$ coul, $\epsilon_n = 6.2 \times 10^{-21}$ $F/Å$, $\lambda = 1$ Å [1] and $E_0 = 0.8$ eV [1], Eq. (8) gives $N_0 = 9.3 \times 10^{20}$ cm^{-3}, which is somewhat higher than the value of 6×10^{-19} cm^{-3} given by Ross and Wallmark.

From Fig. 6, Eq. (58) and the value of N_0, we find that $\omega_0 = 10^{11}$ sec^{-1} which is a bit lower than the value of 10^{13} sec^{-1} calculated from the theoretical expression derived by Dorda and Pulver.

The trap density raises questions as to whether there might be overlap among the wave functions of the stored charge. If this be the case, at least some interfacial areas may exhibit metallic characteristics including lateral conduction, as a consequence of which the simple model of direct tunnelling may need to be modified. Further work is needed to resolve these points.

8. Conclusions

A model for charge transfer in the direct tunnelling mode has been developed with a trap distribution that is uniform, both energetically and spatially. It is shown that this model is a generalization of the models proposed by Dorda and Pulver, and by Ross and Wallmark. The charge transport equation is solved to show that the time dependence is initially linear, then roughly logarithmic and eventually saturation sets in. The operational dependence is found to be quite sensitive to small variations in tunnelling probability, de Broglie wavelength and, particularly, oxide thickness. The switching time varies inversely with trap density and tunnelling probability, exponentially with oxide thickness, and as the exponential of the reciprocal de Broglie wavelength. For high fields, as in typical operation, there is no field dependence within the framework of this model. Experimental evaluation of maximum tunnelling distance, trap density and tunnelling probability gave results in substantive agreement with values in the literature.

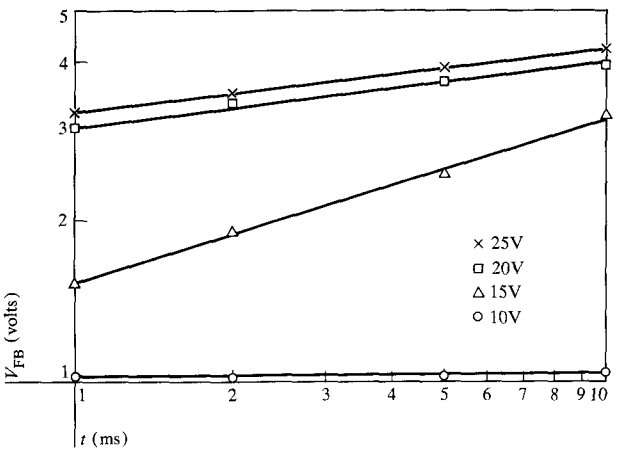


Figure 6 Flatband voltage vs time.

Acknowledgments

It is a pleasure to thank J. J. Chang, who suggested this work, and B. Agusta for their encouragement. Helpful discussions with J. J. Chang, B. V. Keshavan and P. C. Arnett are acknowledged as also are helpful comments by N. J. Chou and P. Balk. I am especially indebted to L. D. Burns who executed the numerical analysis with elegance.

References

1. E. C. Ross and J. T. Wallmark, "Theory of the Switching Behavior of MIS Memory Transistors," *RCA Review* **30,** 366 (1969).
2. D. Frohman-Bentchkowsky and M. Lenzlinger, "Charge Transport and Storage in Metal-Nitride-Oxide-Silicon (MNOS) Structures," *J. Appl. Phys.* **40,** 3307 (1969).
3. G. Dorda and M. Pulver, "Tunnel Mechanism in MNOS Structures," *Phys. Stat. Solidi (a)* **1,** 71 (1970).
4. K. I. Lundström and C. M. Svensson, "Theory of the Thin-Oxide MNOS Memory Transistors," *Electronics Letters* **6,** 645 (1970).
5. K. I. Lundström and C. M. Svensson, "Properties of MNOS Structures, *IEEE Trans. Electron Devices* **ED-19,** 826 (1972).
6. L. G. Carlstedt and C. M. Svenson, *IEEE J. Solid-State Circuits* **SC-7,** 382 (1972).
7. Y. Uchida et al. "A Novel Mode of Write Operation Utilizing Avalanche Tunnel Injection in MNOS Memory Transistors," Paper 13.3, Session 13, Integrated Circuits-3, Nonvolatile Memory Technologies. *IEEE Int'l Electron Devices Meeting*, Washington, D.C., December 5, 1972.

8. L. S. Wei, "Thermal Characterization of Memory Effects in MNOS Capacitors," Paper 13.4, *ibid.*
9. C. Svensson and I. Lundström. "Trap assisted charge injection in MNOS Structures," Paper 13.5, *ibid.*
10. J. R. Cricci, "Random Access MNOS/SOS Nonvolatile Memory," Paper 13.6, *ibid.*
11. P. Balk and F. Stephany, "Charge Storage in MAOS Structures," *NTZ-Nach. Tech. Z.* **10,** 526 (1970).
12. A. V. Ferris-Prabhu, "Time Dependence of Charge Transport in MIS Memory Transistors," *Appl. Phys. Letters* **20,** 149 (1972).
13. A. V. Ferris-Prabhu, "Maximum Tunnelling Distance in MNOS Device Theory," *Phys. Stat. Solidi (a)* **11,** 81 (1972).
14. E. J. M. Kendall, *Phys. Stat. Solidi* **32,** 763 (1969).
15. G. A. Brown et al. "Electrical Characteristics of Silicon Nitride Films Prepared by Silane Ammonia Reaction," *J. Electrochem. Soc.* **115,** 948 (1968).
16. S. M. Sze, "Current Transport and Maximum Dielectric Strength of Silicon Nitride Films," *J. Appl. Phys.* **38,** 2951 (1967).
17. E. J. M. Kendall, "The Conduction Processes in Silicon Nitride," *Can. J. Phys.* **46,** 2509 (1968).
18. *Handbook of Mathematical Tables, National Bureau of Standards (US) Appl. Math. Series* **55** (1964), 228.
19. See, e.g., V. Kourganoff, *Basic Methods in Transfer Problems*, Clarendon Press, Oxford 1952, p. 258.
20. E. C. Ross, et al., "Operational Dependence of the Direct-Tunnelling Mode MNOS Memory Transistor on the SiO_2 Layer Thickness," *RCA Review* **31,** 467 (1970).
21. N. J. Chou et al., "Effect of Insulator Thickness Fluctuations on MNOS Charge Storage Characteristics," *IEEE Trans. Electron Devices* **ED 19,** 2198 (1972).
22. A. M. Goodman, et al., "Optimization of Charge Storage in the MNOS Memory Device," *RCA Review* **31,** 342 (1970).
23. B. H. Yun and P. C. Arnett, "Transient Charge-Transport in MNOS Memory Devices," Paper 13.1, Session 13, Integrated Circuits-3, Non Volatile Memory Technologies, *IEEE International Electron Devices Meeting*, Washington, D.C., December 5, 1972.

Received November 3, 1972

The author is located at the IBM System Products Division Burlington laboratory at Essex Junction, Vermont 05452.