Direct and trap-assisted elastic tunneling through ultrathin gate oxides

F. Jiménez-Molinos, a) F. Gámiz, A. Palma, P. Cartujo, and J. A. López-Villanueva
Departamento de Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, Av. Fuentenueva s/n, 18071 Granada, Spain

(Received 3 July 2001; accepted for publication 23 January 2002)

The direct and assisted-by-trap elastic tunnel current in metal–oxide–semiconductor capacitors with ultrathin gate oxide (1.5–3.6 nm) has been studied. Bardeen’s method has been adapted to obtain the assisted tunnel current, in addition to the direct tunnel current. The dependence of the assisted current on the trap distribution in energy has also been analyzed. This allows us to obtain the trap distribution in energy from experimental current curves. Finally, we have analyzed the role of the image force, the inclusion of which can avoid a barrier height dependence on the oxide thickness.


I. INTRODUCTION

When the channel length is reduced, the gate silicon dioxide must be made thinner to control the short-channel effects and to get a good subthreshold turn-off slope. Taking this into consideration, the oxide thickness must be around 3.0 nm in 0.1 μm metal–oxide–semiconductor (MOS) technology. The tunnel current in such conditions is very important (it may exceed the maximum tolerable standby power when the oxide is thinner than 2.0 nm) and affects the device characteristics, as has been shown experimentally. Therefore, it is obvious that the performance of any simulator of devices with these dimensions should consider the contribution of the tunnel current. Although it will no longer be possible to further scale down SiO2 based gate dielectrics because of tunneling leakages, the near-term gate dielectrics require their fabrication and use until suitable alternative high dielectric constant material with good stability and interface characteristics can be found.

In order to model the tunnel current, the envelope function and effective mass approximations are usually assumed. Despite their good behavior, there is no rigorous theoretical basis for their use. In a recent paper, Staedle et al. tested these approximations against a microscopic model based on a semiempirical tight-binding scattering method. They concluded that the effective mass approximation qualitatively reproduces the energy and voltage dependence of the transmission coefficients if an energy-dependent effective mass is used. Nevertheless, the real value is overestimated. The fitting parameters usually used in these models are presented.

The article is organized as follows. Section II shows the theoretical model used to calculate the tunnel current. Section III presents some simulations showing the influence of the image force and the trap distribution in energy. A comparison between experimental data and simulation results is given in Sec. IV. Finally, the main conclusions of this work are presented.

II. THEORETICAL MODEL

In this article, we examine tunnel transitions from the p-Si(100) substrate to the n+ polysilicon gate in a MOS. A positively biased gate is assumed and therefore an inversion layer is formed. As is widely known, the potential well that confines the electrons near the Si–SiO2 interface is very narrow. In consequence, quantum effects are very important and we should take into consideration the electron energy quantization in the transverse direction because of the potential energy.

\[ V(z) = -e\phi(z) + E_{im}(z) + E_{C0}(z), \]

where \( z \) is the coordinate along the transverse direction to the interface Si–SiO2, \( \phi(z) \) is the electrostatic potential solution.

a)Electronic mail: jmolinos@ugr.es
of Poisson’s equation, \( E_{im}(z) \) is the image energy (if included) and \( E_{C0}(z) \) takes account of the conduction band discontinuity:

\[
E_{C0}(z) = \begin{cases} 
0 & \text{in Si} \\
\phi_B & \text{in } \text{SiO}_2.
\end{cases}
\]

Because of the energy quantization, Poisson and Schrödinger equations have to be self-consistently solved. When the image effect is considered, its calculation is performed according to Ref. 20, where the image energy is calculated for a MOS structure:

\[
E_{im}(z) = \left( \frac{-q^2}{16 \pi \epsilon_S} \right) \sum_{n=0}^{\infty} \left( k_1 k_2 \right)^n \frac{-k_1}{|z + n \tau|} + \frac{k_2}{|z + (n+1) \tau|},
\]

within the semiconductor and

\[
E_{im}(z) = \left( \frac{-q^2}{16 \pi \epsilon_{ox}} \right) \sum_{n=0}^{\infty} (k_1 k_2) \times \left( \frac{k_1}{|z + n \tau|} + \frac{k_2}{(n+1) \tau - |z|} + \frac{2 k_2}{(n+1) \tau} \right),
\]

within the oxide. The Si–SiO\(_2\) interface is assumed to be at the z-axis origin. \( \epsilon_{ox} \) and \( \epsilon_S \) are the oxide and silicon dielectric constants, respectively; \( q \) is the electron charge; \( \tau \) is the oxide thickness, and \( k_1 \) and \( k_2 \) are the reflection coefficients:

\[
k_1 = \frac{\epsilon_{ox} - \epsilon_S}{\epsilon_{ox} + \epsilon_S}, \]

\[
k_2 = -1.
\]

The gate material is assumed to be metal-like. In practice, it is sufficient to take the sum from \( n = 0 \) to \( n = 11 \). To avoid the divergence of the image term at the interfaces, we use Eqs. (2) and (3) to calculate the image energy in regions further than 1 Å from these divergence points. Between these points, we carry out an interpolation in the total potential energy \( V(z) \).

For the calculation of the tunnel current, we must obtain the transition probabilities for direct tunneling of an electron and for trapping and detrapping processes (assisted tunneling). We have followed Bardeen’s method\(^{12-15}\) for both calculations, as explained in the next subsections. The effective mass in the oxide has been taken as energy dependent, following the analytical expression recently reported by M Städle et al.\(^5\)

\[
\frac{m_{ox}(e)}{m_{ox}(0)} = 1 + \sum_{i=1}^{3} a_i e^i,
\]

with \( a_1 = -0.178 \text{ eV}^{-1} \), \( a_2 = 0.025 \text{ eV}^{-2} \) and \( a_3 = -0.0023 \text{ eV}^{-3} \). Finally, \( \epsilon \) is the energy, measured from the bottom of the oxide conduction band toward the valence band.

According to the results of Ref. 5, the calculations of tunnel current based on the envelope function and effective mass approximations with the energy-dependent effective mass given by Eq. (6) reproduce the correct slope of the transmission coefficients versus voltage or electron energy curves. However, to reproduce experimental data, adjustable parameters [such as \( \tau \), \( \phi_B \), or \( m_{ox}(0) \)] must be taken.

In the present work, we have used a single value of \( m_{ox}(0) \) for all the samples (equal to 0.6\( \text{m}_0 \)), the value usually taken with the Franz-type dispersion relationship. For the barrier height, as it will be shown in Sec. 4, we have used several values depending on whether the image force is included or whether we take into account only the current from the two-fold-degenerated valleys. But in any case, no fitting parameters are necessary to reproduce the experimental data of samples with different oxide thicknesses (except a weak dependence of \( \phi_B \) if the image force is not included): with the same values of \( m_{ox}(0) \) and \( \phi_B \) all the curves are reproduced well, as shown next.

### A. Direct tunnel current

The substrate is assumed to be p-Si(100) and the gate is modeled as \( n^+\)-Si(100). The total tunnel current density is given by

\[
J = J_{1-2} + J_{2-1} = \epsilon \sum_{E_{gj}} \int_{E_1} dE \rho(E) \rho_1(E_g) \rho_2(E_{g\mu}) f_{1}(E) - f_{2}(E),
\]

where subscripts 1 and 2 refer to the silicon and polysilicon sides, respectively. The quantization of the perpendicular energy and the corresponding formation of subbands (labeled by \( \mu \)) in the inversion layer have been taken into consideration. \( \omega(E) \) is the probability per time unit of transition of an electron with energy \( E \) (measured from the bottom of the silicon conduction band) from one side of the barrier to the other (irrespective of which side the electron leaves from). \( \rho_1(E_g) \) is the bidimensional density of states with parallel energy \( E_g \):

\[
\rho_1(E_g) = g_v \frac{\sqrt{m_x m_y}}{\pi \hbar^2},
\]

where \( g_v \) is the valley degeneration. \( \rho_2(E_{g\mu}) \) is the unidimensional density of states in the gate with perpendicular energy \( E_{g\mu} \) (the parallel moment \( K_z \) is conserved during the process, as shown in Appendix A):

\[
\rho_2(E_{g\mu}) = \frac{L_z m_z}{2 \pi \hbar^2 K_z}.
\]

Spin degeneration is not included because the spin is conserved during the tunneling. \( L_z \) is a length along the \( z \) axis and will be cancelled out. The functions \( f_1(E) \) and \( f_2(E) \) in Eq. (7) are the Fermi–Dirac functions. The probability \( \omega(E) \) is given by

\[
\omega(E) = \frac{2 \pi}{\hbar} \frac{\frac{\hbar^2}{2 m_{ox}(V(z_{max}) - E_{g\mu})}^2}{\left( \frac{\partial \zeta_x}{\partial z} - \frac{\partial \zeta_y}{\partial z} \right)^2 z_{z_{max}}^2 \delta_{K_z K_z}},
\]

Downloaded 28 Apr 2005 to 150.214.68.115. Redistribution subject to AIP license or copyright, see http://jap.aip.org/jap/copyright.jsp
Detailed calculations of $\omega(E)$ are shown in Appendix A, as well as the meaning of the symbols. Finally, by substituting Eqs. (10) into (7) and operating, we get

$$J = \frac{eKT}{4\pi\hbar} \sum_{\mu} \frac{g_0 \sqrt{m_\mu m_e m_i}}{m_\mu \epsilon(z_{\max} - E_{\mu})} \zeta \times \left[ \frac{\partial \zeta}{\partial z} - \frac{\partial \zeta}{\partial z} \right] \ln \left( \frac{e^{(-E_z + E_T)/KT} + 1}{e^{(-E_z + E_T - qV_g)/\hbar KT} + 1} \right).$$

(11)

We have obtained a formula for the tunnel current density $J$, but the penetration of the wave functions into the oxide must still be calculated. This can be done numerically, solving the Poisson and Schrödinger equations or by several analytical approximations. We have followed the first method because it is more general and can be used for different barrier shapes and expressions of $m_\mu(E)$. Therefore, some parameters, such as $m_\mu$ and $\phi_B$, appear in the direct current density formula (11) but they also influence $J$ by changing the penetration of the wave functions within the oxide.

In the calculation of the tunnel current, we have taken into account only the direct tunnel current from the conduction band. The tunnel current from the valence band is at least two orders of magnitude smaller than that of the conduction band, because the barrier is considerably higher, and therefore it was neglected.

### B. Elastic trap-assisted tunnel current

We have considered trap-assisted tunneling as a two-step process. Firstly, electrons from the cathode are trapped and then they are detrapped to the anode. If we neglect the tunnel current in the opposite direction to the electrical field in the oxide, the change ratio of the concentration of occupied traps in position $z$ is

$$\frac{dn_\mu(z,E)}{dt} = N_T(z,E)[1 - f_\mu(z,E)]\tau_e^{-1}(z,E) - N_T(z,E)f_\mu(z,E)\tau_e^{-1}(z,E)$$

$$= R(z,E) - R(z,E),$$

(12)

where $E$ is the energetic position of the traps (that is $E = V(z) - E_T$, where $E_T$ is the trap deepness or distance between the bottom of the oxide conduction band and the energy level of the traps), $N_T(z,E)$ is the concentration of traps with energy $E$ at position $z$, $f_\mu(z,E)$ is their occupancy function and $\tau_e(z,E)$ and $\tau_c(z,E)$ are the time constants for the capture and emission processes, respectively. Positive bias has generally been assumed (electrons go to the $n^+$-polysilicon gate from the $p$ substrate in inversion or $n$ substrate in accumulation). Therefore, the capture time is given by

$$\tau_e^{-1}(z,E) = \sum_{\mu} \tau_{e,\mu}(z,E),$$

(13)

where the summation is over all subbands ($\tau_{e,\mu}$ is the capture time for electrons from the $\mu$ subband). In steady state,

$$\frac{dn_T(z,E)}{dt} = 0,$$

(14)

and

$$R(z,E) = R(z,E) = R.$$

(15)

Therefore,

$$R(z,E) = \frac{N_T(z,E)}{\tau_c(z,E) + \tau_e(z,E)},$$

(16)

where $R(z,E)$ is the number of transitions per unit time assisted by the traps with energy $E$ at position $z$. The total assisted tunnel current density is then given by:

$$J_T = q \int_0^\infty dz \int_{E_T}^{E_e} \frac{N_T(z,E)}{\tau_c(z,E) + \tau_e(z,E)} dE,$$

(17)

where $q$ is the electron charge and $E_0'$ is the energy of the ground subband.

The time constants are calculated from:

$$\tau_c(z,E) = \sum_{\mu} D_{c,2D}(E) f_\mu(z,E) W_c(z,E),$$

(18)

$$\tau_e(z,E) = \int_{E_c}^E D_{a,2D}(E) D_{a,1D}(E) \left[ 1 - f_\mu(E) \right] W_a(z,E) dE,$$

(19)

where $D_{c,2D}(E)$ (or $D_{a,2D}(E)$) is the bidimensional density of states in the cathode (anode), $D_{a,1D}(E_c)$ is the one-dimensional (1D) density of states in the anode, $E_c$ is the anode conduction band edge and $f_\mu(E)$ is the Fermi–Dirac function in the cathode (anode). For the emission process, final states with different transversal energy, $E_z$, are possible if the total energy ($E_z + E_1$, where $E_1$ is the parallel energy of the final state) is equal to the trap energy level. $W_c(z,z)$ are the probabilities per time unit of capture and emission, respectively, of an electron by the trap and are given by

$$W_c(z,z) = \frac{2\pi}{\hbar} \sqrt{\frac{\zeta(z)}{\zeta(z + 1/2)}} \left| \int_{z - 1/2}^{z + 1/2} \xi(z') dz' \right|^2.$$ 

(20)

Detailed calculations of the probabilities $W_c(z,z)$ are shown in Appendix B, as well as the meaning of the symbols. As can be seen, these probabilities depend on the transversal energy, $E_z$, because of the different penetration of the wave function into the barrier.

If the initial state is not bound [three-dimensional (3D) electron gas], the capture times should be calculated according with the following expression:

$$\tau_e^{-1}(z,E) = \int_{E_c}^E D_{c,2D}(E) D_{c,1D}(E) f_\mu(E) W_a(z,E) dE.$$ 

(21)

The symbols have analogous meaning as in expression (19), but corresponding to the cathode.
III. SIMULATIONS

Before making a comparison with the experimental results of direct and assisted tunnel current density in MOS capacitors, let us study its dependence on some parameters.

A. Image force and barrier height

There has been some controversy about the inclusion of the classical image energy in the calculation of tunnel current. According to some authors,\textsuperscript{17,22,23} the classical image energy overestimates the actual image force in tunneling problems, and better results are obtained by omitting it from calculations. However, other authors defend its inclusion: in Ref. \textsuperscript{24}, the classical image potential is said to be necessary to describe the barrier-width dependence and the absolute value of the vacuum tunnel current. In Ref. \textsuperscript{8}, the image force is included to avoid a thickness-dependent tunneling mass in the Fowler–Nordheim regime.

Inclusion of the image force is known to lower and round the barrier. The question arises whether it is possible to reproduce the image force effect on the tunnel current curves with only a change in the barrier height $\phi_B$, the value of which is usually used as a fitting parameter in tunneling problems. To determine this, we performed the simulations shown in Fig. 1 with three different oxide thicknesses. In both cases, it can be seen that the curves calculated including the image force are almost exactly reproduced by other simulations that exclude it but in which the value of $\phi_B$ is reduced. Therefore, in principle, we could avoid the image force, especially if $\phi_B$ is used as a fitting parameter. However, the reduction of $\phi_B$ depends on the oxide thickness. Thus the inclusion of the image force could explain a reduction of the barrier height when oxide thickness is decreased. This dependence of $\phi_B$ on $t_{\text{ox}}$ has been described by Refs. \textsuperscript{11} and \textsuperscript{25} and will also be shown in the following section.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1}
\caption{Image force and barrier height influence on the simulated direct tunnel current density curves $[N_{\text{sub}}=5\times10^{17}\text{ cm}^{-3}(P); N_{\text{poly}}=5\times10^{20}\text{ cm}^{-3}(N)]$. In all of these simulations, $m_{\text{ox}}(0)=0.6m_0$.}
\end{figure}

The above is illustrated in Figures 1 and 2. The image energy lowers $V(z)$ near the interfaces, but not far from them. Therefore, the thinner the oxide, the more important the lowering effect. In fact, in very thin oxides the two barriers (with and without the inclusion of the image force) can be different even in the middle of the oxide. Therefore, if the oxide thickness decreases, $\phi_B$ should be lowered by a greater amount. Figure 1 shows that reducing the barrier in the cases of $t_{\text{ox}}=2.00\text{ nm}$ and $t_{\text{ox}}=1.60\text{ nm}$ by the same amount as in the thicker oxide moves the curve below the curve that includes the image force.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2}
\caption{Image force and barrier height influence on the well $V(z)$ and the electron density $n(z)$. The device parameters are the same as in Fig. 1. The gate voltage is $V_g(z)=2.0\text{ V}$. Solid line: $\phi_B=3.15\text{ eV}$, without $E_{\text{im}}$. Dashed line: $\phi_B=2.9\text{ eV}$, without $E_{\text{im}}$. Dashed and dotted line: $\phi_B=2.8\text{ eV}$, without $E_{\text{im}}$. Dotted line: $\phi_B=3.15\text{ eV}$, with $E_{\text{im}}$.}
\end{figure}

B. Trap distribution in energy

In this subsection, we analyze the influence of the trap distribution in energy on the assisted tunnel current curves. The distribution can be discrete or continuous. In the latter case, the distribution in energy has been taken as Gaussian, following

$$N_T(E_T,z)=\begin{cases} N_0e^{-[(E_T-E_{\text{m}})/\Delta E]^2} & \text{if } E_T \in [E_{\text{m}},E_{\text{m}}] \\ 0 & \text{if } E_T \notin [E_{\text{m}},E_{\text{m}}] \end{cases}$$

(22)

\[5119\] J. Appl. Phys., Vol. 91, No. 8, 15 April 2002

Jiménez-Molinos et al.

Downloaded 28 Apr 2005 to 150.214.68.115. Redistribution subject to AIP license or copyright, see http://jap.aip.org/jap/copyright.jsp
where $E_{\text{min}}$ is the minimum distance in energy between the traps and the bottom of the oxide conduction band and $E_{\text{max}}$ corresponds to the deepest traps. The average of the distribution is $E_m = 0.5(E_{\text{max}} + E_{\text{min}})$ and the width is controlled by the parameter $\Delta E$. In order for the concentration at the edge of the distribution to be approximately a tenth that in the middle, it has been taken as $\Delta E = 0.65(E_{\text{max}} - E_m)$.

Figure 3 shows the results obtained with several distributions. In all of them, $t_{\text{ox}} = 2.8$ nm, $N_0 = 10^{17}$ cm$^{-3}$, $\phi_B = 3.15$ eV, the trap distribution in energy does not depend on the position and the image energy has not been included. In the case of a single-energy level (discrete distribution), we can observe a peak that corresponds to the alignment between the baricentric traps and the most populated subbands. This peak disappears when a continuous distribution is assumed. As shown in Fig. 3, if the center of the distribution is around $\phi_B$ (basis case in Fig. 3), the assisted tunnel current is noticeable at low biases. However, at medium biases it decreases rapidly because the traps are below the subbands, unless the distribution is wider. If the distribution is less deep, the curves are displaced to higher biases.

Inversely, the shape of the assisted tunnel current curves can give us important information about the traps as shown in the next section: only with the parameters shown in Table I, the trap-assisted tunnel current reproduces the experimental data not fitted by direct tunnel current.

### TABLE I. Parameters of the trap distribution in energy used to fit the experimental data in Fig. 5.

<table>
<thead>
<tr>
<th>$t_{\text{ox}}$ (nm)</th>
<th>$t_{\text{ox}}$ (nm)</th>
<th>$t_{\text{ox}}$ (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_0$ (cm$^{-3}$ eV$^{-3}$)</td>
<td>$4 \times 10^{14}$</td>
<td>$1 \times 10^{17}$</td>
</tr>
<tr>
<td>$E_{\text{min}}$ (eV)</td>
<td>2.0</td>
<td>2.2</td>
</tr>
<tr>
<td>$E_{\text{max}}$ (eV)</td>
<td>3.0</td>
<td>3.2</td>
</tr>
</tbody>
</table>

C. Comparison with an one-step model

Figure 4 shows a comparison with one-step model, based on the transfer matrix method$^{26,27}$ and the expressions given in Ref. 26. In the simulations, negative gate voltage has been assumed because the compared model assumes incident plane waves (3D electronic gas). As can be seen, the results of both models are very closed.

The model presented in this work can be also employed to calculate tunnel current from a two-dimensional (2D) electronic gas, with arbitrary barrier shape (not only trapezoidal) and $m_{\text{ox}}(E)$ dependence in an easy way. However, obviously, it consumes more computer time than an analytical model.

IV. COMPARISON WITH EXPERIMENTAL RESULTS

Experimental results from Refs. 2 and 7 and our simulated curves are shown in Fig. 5. We have used expression (6) reported in Ref. 5 for the energy-dependent effective mass in the oxide. The value for the effective mass at the bottom of the conduction mass has been taken as 0.6$m_0$ (equal to the value usually taken in the Franz-type dispersion relationship). For the barrier height, the value 3.5 eV has been assumed. Image energy has been included. No fitting parameters have been used for the direct tunnel current calculation (all the curves are calculated with the same values of $m_{\text{ox}}(0)$ and $\phi_B$). As can be seen, the experimental data corresponding to the oxides thinner than 2.91 nm are well reproduced taking into account only the direct tunnel current, except for low biases. However, for the thickest oxides, the direct tunnel curves are below the experimental data. The difference is attributed to elastic tunneling assisted by native traps.$^{28,29}$ We have included its contribution as in the model shown in Sec. II. Uniform spatial distribution of traps has been assumed. The distribution in energy has been taken as continuous with a Gaussian shape, following expression.
For good fitting of the experimental data, we had to take the values shown in Table I. As can be seen, the energy level of the traps is similar in the three samples analyzed (with different oxide thicknesses), but the maximum concentration changes by one order of magnitude.

These experimental data can also be reproduced without including image force in the calculation of the well $V(z)$ [expression (1)], although a weak dependence of the barrier height on the oxide thickness is observed: in this case, the value 3.2 eV has been assumed for the barrier height, $\phi_B$, but a slightly better result is obtained using $\phi_B = 3.1$ eV for the oxide with $t_{ox} = 2.19$ nm and $\phi_B = 3.0$ eV for the oxide with $t_{ox} = 1.55$ nm. On the other hand, the parameters of the trap distributions are the same, but shifted 0.3 eV to the bottom of the oxide conduction band. In this way, from the point of view of the alignment of the traps with the bottom of the silicon conduction band in the neutral zone, the trap distribution is approximately the same as with $\phi_B = 3.5$ eV with the inclusion of the image force. Therefore, the importance of the traps is mainly given by their concentration and energy relative to the bottom of the silicon conduction band in the neutral zone.

The discrepancy at low biases for the samples with oxide thinner than 2.5 nm can be avoided if we take into account only the current of the electrons that belong to the two-fold-degenerated valleys, as shown in Fig. 6. We can justify the fact of including only this contribution if we take into consideration one of the results of Ref. 5: the parallel moment of the electrons is conserved if the oxide thickness is thinner than a certain critical value between 1 and 3 nm. In the case of conservation of the parallel moment, the electrons belonging to the four-fold-degenerated valleys “see” a greater barrier height$^{5,30}$ (as illustrated in Fig. 7) and, therefore, their contribution could be neglected.

On the other hand, as in the case of including the current from both kinds of valleys, we have had to assume an oxide thickness-dependent barrier height if the image energy is not included, as illustrated in Fig. 6.

The reason for the thickness dependence might be found in the three following effects: (a) the thinner the oxide, the greater the influence of the image energy on the barrier shape. As a consequence, the barrier height must be made lower when the oxide thickness decreases if the image force is not taken into account, as stated in the previous section. (b) As mentioned herein, the barrier height is thickness dependent for oxides thinner than the critical value. Below the...
critical thickness, the parallel moment is conserved and, therefore, the electrons belonging to a different kind of valley see a different barrier. Above the critical thickness, the parallel moment is not conserved and the electrons see the same barrier. Between the two situations, the barrier is thickness dependent. (c) The nonstoichiometric transition region in the interfaces could induce a thickness-dependent barrier height change, as suggested by some authors. To avoid this thickness dependence of the barrier height, we have included the image force in the calculation of the subbands and transmission probabilities. We have found that it is possible to fit the experimental data of the oxides in the range 1.55–2.56 nm using the same barrier height (\(\phi_B = 3.25\) eV) and no fitting parameters if the image force is taken into account. These curves are not reproduced, to avoid repetition. The image energy has previously been used to avoid thickness dependence of the oxide effective mass and barrier height.

V. CONCLUSIONS

Direct and trap-assisted elastic tunnel current through gate oxides has been studied. We have developed a model, based on Bardeen’s method, for the calculation of the trap-assisted elastic tunnel current, in addition to the usual direct tunnel current. We have included the analytical energy-dependent oxide effective mass recently proposed by Städele et al. This avoids the use of a thickness-dependent oxide effective mass. We have also analyzed the dependence of the assisted tunnel current on the trap distribution in energy. Inversely, the trap distribution could be determined from experimental assisted tunnel current curves.

We have fitted experimental data of tunnel current through gate oxides in MOS devices. For oxide thicknesses over 2.60 nm, it is necessary to take into account the elastic tunneling assisted by traps. We have assumed a Gaussian trap distribution in energy, uniform at all the points in the oxide. However, for the thinnest oxides, the direct tunnel current is higher than the assisted tunnel current and reproduces the experimental curves well, except with low biases. To get a good fit at these points, we need to assume that the main contribution to the tunnel current is from the two-fold-degenerated valleys. This assumption is supported by the fact that the parallel moment of the electrons is conserved if the oxide thickness is thinner than a certain critical value between 1 and 3 nm.

Furthermore, the inclusion of the image energy is necessary to avoid a weak dependence of the barrier height, \(\phi_B\), on the oxide thickness. This does not imply that all the thickness dependence of the barrier is caused by the image force, but the real cause could be a mixture of several effects. The values used for \(\phi_B\) are the following (same for all the samples, it is not a fitting parameter): (i) if only the two-fold-degenerated valleys contribute to the current and image force is included then \(\phi_B = 3.25\) eV. (ii) If we take into account the contribution of both kinds of valleys and the image force, then \(\phi_B = 3.5\) eV. (iii) Without including image force, we can reproduce the experimental data approximately with \(\phi_B \approx 3.0\) eV if only the two-fold degenerated valleys contribute to the current and with \(\phi_B \approx 3.2\) eV if all the valleys contribute; however, as said above, a weak dependence of \(\phi_B\) on the oxide thickness is observed if the image force is not included.

Neither the value of \(\phi_B\) nor the inclusion of the image energy affects the trap distribution in energy, which is shifted from the top of the oxide conduction band approximately by the same amount as \(\phi_B\) changes. In this way, the trap distribution seen from the bottom of the silicon conduction band is the same.

ACKNOWLEDGMENT

This work has been carried out within the framework of Research Project No. PB97-0815, supported by the Spanish Government.

APPENDIX A: DIRECT TUNNEL TRANSITION

In this Appendix, the calculation of the probability per time unit of direct transition of an electron with energy \(E\) from one side of the barrier to the other \([\omega(E)]\) following Bardeen’s method is detailed. We calculate the probability \(\omega\) in the context of the envelope function and effective mass approximations. The envelope function is given by the solution of the Schrödinger equation with the Hamiltonian \(H = T + V(z)\), where \(V(z)\) is given by Eq. (1). We define the following Hamiltonians

\[
H_1 = \begin{cases} 
T + V(z) & \text{if } z < z_{\text{max}} \\
T + V_{\text{max}} & \text{if } z > z_{\text{max}},
\end{cases}
\]

\[
H_2 = \begin{cases} 
T + V_{\text{max}} & \text{if } z < z_{\text{max}} \\
T + V(z) & \text{if } z > z_{\text{max}},
\end{cases}
\]

where \(z_{\text{max}}\) is the point on the \(z\) axis where \(V(z)\) reaches its maximum \(V_{\text{max}}\). Then, we can define \(H = H_1 + H'\) with:

\[
H' = \begin{cases} 
0 & \text{if } z < z_{\text{max}} \\
H_2 - H_1 & \text{if } z > z_{\text{max}}.
\end{cases}
\]

Using the Fermi golden rule, \(\omega\) becomes

\[
\omega(E) = \frac{2\pi}{\hbar} \left| \int_{z_{\text{max}}}^{\infty} dz \int_A d^2 R \Psi_2^*(r) (H_2 - H_1) \Psi_1(r) \right|^2,
\]

where \(\Psi_1(r)\) and \(\Psi_2(r)\) are the eigenvalues of \(H_1\) and \(H_2\), respectively, and are given by \(\Psi_1(r, K_\parallel) = 1/\sqrt{A} \xi_1(z) e^{i K_\parallel \cdot R}\) and \(\Psi_2(r, K_\parallel) = 1/\sqrt{A} \xi_2(z) e^{i K_\parallel \cdot R}\). This form of the envelope function implies the conservation of parallel moment \(K_\parallel\), because the Hamiltonian \(H'\) considered here does not include any term which connects states of \(H_1\) and \(H_2\) with different parallel moments.

Before going on, we make \(H_2\) operate over \(\Psi_2^*\) instead of \(\Psi_1^*\):

\[
\int_{z_{\text{max}}}^{\infty} dz \int_A d^2 R \Psi_2^* H_2 \Psi_1 = \int_{z_{\text{max}}}^{\infty} dz \int_A d^2 R \Psi_1 H_2 \Psi_2^{*} - i\hbar \int_A d^2 R \frac{d}{dR} \Psi_1(r) R_{12}^r(r),
\]
where

\[ J_{12}^\Psi (R) = \frac{\hbar}{2i m_\text{ox}(V(z_\text{max}) - E)} \left[ \Psi_1 \frac{\partial \Psi_2}{\partial z} - \Psi_2 \frac{\partial \Psi_1}{\partial z} \right]_{z = z_\text{max}} \]  

(27)

as can be shown integrating \( \int_{z_\text{max}}^\infty dz \Psi_2^* T \Psi_1 \) by parts twice. Performing Eq. (25), taking into consideration that \( \Psi_1 \) is an eigenfunction of \( H_1 \) with the same eigenvalue \( E \) as \( \Psi_2 \) is an eigenfunction of \( H_2 \), we can conclude:

\[ \omega(E) = \frac{2 \pi}{\hbar} \left( \frac{\hbar^2}{2m_\text{ox}(V(z_\text{max}) - E)} \right)^2 \times \left[ \frac{\partial \xi_2}{\partial z} - \frac{\partial \xi_1}{\partial z} \right]_{z = z_\text{max}}^2 \delta \kappa_k, \kappa'_k. \]  

(28)

Finally, substituting Eqs. (28) into (7) and operating, we get Eq. (11). The length \( L_z \) in expression (9) has been cancelled out with the normalization of the wave function \( \xi_2 \), calculated by solving the Schrödinger equation with the Hamiltonian \( H_2 \).

**APPENDIX B: TRAP-ASSISTED TUNNEL TRANSITION**

In this Appendix, the calculation of the probability per time unit of trapping or detrapping of an electron by a trap \([W_c(a)(z,E)]\) following Bardeen's method\(^\text{12}\) is detailed. We will focus on the trapping transition from the substrate. The detrapping transition is calculated in the same way.

The trap is assumed to be a cubic box with volume \( V_T \) and side length:

\[ x_T = y_T = z_T = \frac{4 \pi \hbar}{\sqrt{2m_\text{ox}(0)E_T}}. \]  

(29)

and whose center is at \( z_0 \). The total Hamiltonian is \( H_T = H_1 + H' \), where

\[ H_T = \begin{cases} 
T + V(z) - E_T & \text{if } z \in \left[ z_0 - \frac{z_T}{2}, z_0 + \frac{z_T}{2} \right], \\
H_1 & \text{if } z \notin \left[ z_0 - \frac{z_T}{2}, z_0 + \frac{z_T}{2} \right].
\end{cases} \]  

(30a)

\[ H_1 = \begin{cases} 
T + V(z) & \text{if } z < t_\text{ox}, \\
T + V(z = t_\text{ox}) & \text{if } z > t_\text{ox}.
\end{cases} \]  

(30b)

\[ H' = \begin{cases} 
-E_T & \text{if } z \in \left[ z_0 - \frac{z_T}{2}, z_0 + \frac{z_T}{2} \right], \\
0 & \text{if } z \notin \left[ z_0 - \frac{z_T}{2}, z_0 + \frac{z_T}{2} \right].
\end{cases} \]  

(30c)

The Hamiltonian \( H' \) is now, in the trap region, the difference between the Hamiltonian \( H_T \) and the Hamiltonian \( H_1 \), whose eigenstates are the initial or detrapped states. The origin of the \( z \) axis is on the Si–SiO\(_2\) interface, whereas the SiO\(_2\)-polysilicon interface is at \( z = t_\text{ox} \). The energy level of the trap is \( E_T \), measured from the bottom of the oxide conduction band.

According to Fermi’s golden rule, the probability of transition per time unit between the initial state \([\langle i \rangle]\) and the trapped state \([\langle f \rangle]\) is:

\[ W_c(z,E) = \frac{2 \pi}{\hbar} |\langle f | H' | i \rangle|^2 \delta(E - E_T). \]  

(31)

where \( E_T \) is the energy of the trap at position \( z \).

\[ E_T = V(z) - E_T. \]  

(32)

The matrix element \([\langle f | H' | i \rangle]\) is calculated from

\[ \langle f | H' | i \rangle = \int \Psi_f^*(r) H'(r) \Psi_i(r) dr, \]  

(33)

with

\[ \Psi_f(r) = \begin{cases} 
\frac{1}{\sqrt{V_T}} & \text{if } \left\{ \begin{array}{l} x \leq \frac{x_T}{2} \\
y \leq \frac{y_T}{2} \\
z \leq \frac{z_T}{2} \end{array} \right., \\
0 & \text{elsewhere}
\end{cases} \]  

(34)

\[ \Psi_i(r) = \frac{1}{\sqrt{A}} e^{i(k_y y + k_z z)} \xi_1(z), \]  

(35)

where \( A \) is the device area and the wave function \( \xi_1(z) \) is normalized over a length \( L_z \) (if the initial state is not bound). Then, we substitute Eqs. (34) and (35) into Eq. (33) and take into account the next approximation:

\[ \int_{x_0}^{x_0 + x_T^2} \int_{y_0 - y_T^2}^{y_0 + y_T^2} \int_{z_0 - z_T^2}^{z_0 + z_T^2} e^{i k_x x_0} e^{i k_y y_0} \sin \left( \frac{k_y y}{2} \right) \xi_1(z) \xi_1(z') dz dz', \]  

(36)

based on the fact that \( k_x x_0 \ll 1 \), with the usual values of \( E_T \) and temperature. Removing the phase, we obtain:

\[ \langle f | H' | i \rangle = \frac{\sqrt{z_T E_T}}{\sqrt{A}} \int_{z_0 - z_T^2}^{z_0 + z_T^2} \xi_1(z') dz'. \]  

(37)

And, finally, introducing Eq. (37) into Eq. (31), we get Eq. (20). The area \( A \) is cancelled out by the corresponding term in the state density expressions.