Gate Tunnel Current in an MOS Transistor

BOGDAN MAJKUSIAK

Abstract—A theoretical description of gate tunnel current in an MOS transistor is proposed, and the results of calculations for the case of an n-channel MOSFET with extremely thin gate oxides are given. A comparison of the gate tunnel current with the drain current is made.

I. INTRODUCTION

PROGRESS in MOS/VLSI technology is accomplished by a constant trend to decrease the gate oxide thickness. This serves two main purposes: 1) to increase the transistor transconductance, which in turn improves the circuit speed, and 2) to avoid short-channel effects [1]. Gate oxides not much above 10 nm thick are used in MOS VLSI circuits already and MOSFET's with ultrathin oxides as thin as 2.5 nm [2] are investigated experimentally in laboratories.

A natural limitation of this trend can result from the tunneling of carriers through the gate oxide. Therefore, it seems useful to consider the influence of gate tunnel current on MOSFET performance. An experimental transistor with a 2.5-nm oxide [2] has exhibited a gate current about nine orders of magnitude lower than the drain current. In order to consider transistors with even thinner oxides, a theoretical description of the gate tunnel current in the MOSFET must be created. A direct transposition of the dc model of the MOS tunnel diode is not sufficient in a general case due to the fact that the band diagram of the semiconductor surface region changes along the source-drain distance in case of non-zero drain-source voltage.

It is the aim of this paper to propose such a description and to compare theoretically the gate tunnel current with the drain current in MOSFET's with extremely thin gate oxides. For practical verification of the model, results will be compared with experimental data given in [2].

II. THEORY

A. Components of the Gate Tunnel Current

Fig. 1 shows the flow paths of the gate tunnel current components, and the energy-band diagram in Fig. 2 illustrates their physical mechanism. An n-channel MOSFET with an n⁺-polysilicon gate is the example considered.

When the gate is biased positively with respect to the substrate, there are two components of the gate tunnel current \( I_G \):

1) the current of minority carriers (electrons) tunneling from the semiconductor substrate to the gate, i.e.,
   \[ I_{G1} \]
2) the current of majority carriers (holes) tunneling from the gate to the semiconductor substrate. In fact, it is a current of electrons tunneling from the semiconductor valence band to free energy states in the gate electrode, i.e., the gate-semiconductor valence band current \( I_{G2} \).

Both components can be distinguished experimentally by measuring the gate and substrate currents at the zero drain-source voltage, i.e., taking advantage of the measurement idea used in [3] for a diode structure. Holes occurring at the surface of the semiconductor are swept mainly by the transverse electric field toward the substrate and can dominate the substrate current provided that the contribution from holes generated thermally or by impact ionization processes in the semiconductor surface region is negligible. Such a situation takes place just when \( V_{DS} = 0 \); otherwise, the substrate current can be dominated by holes generated by channel hot electrons in impact ionization processes. Since simultaneously the gate tunnel...
current at $V_{DS} = 0$ can be modeled as in a uniform MOS capacitor structure, this case of transistor biasing is the most convenient for verification of the theoretical description of gate tunnel current components.

Difficulties in modeling the gate tunnel current appear when $V_{DS} > 0$. In this case, the energy-band diagram of the MOS system changes along the source–drain distance and both components of the gate tunnel currents must be calculated by integration of their densities along the channel.

For the conduction band tunnel current we have

$$I_{GC} = W \int_{\xi_s}^{\xi_e} J_{GC}(\xi) \, d\xi$$

(1)

where $W$ is the channel width, $L$ is the channel length, and $\xi_s = (E_F - E_{Fp})/kT$ is the normalized split of the electron quasi-Fermi level $E_{Fp}$ from the bulk Fermi level $E_F$. $\xi_e$ changes from the normalized source–bulk voltage $v_{SB}$ at the source end of the channel to the normalized drain–bulk voltage $(v_{SB} + v_{DS})$ at the drain end of the channel. $u_i$ is the normalized surface potential and for a given value of $\xi_e$ results from a solution of the equation

$$V_{GS} + V_{SB} - V_{FB} = \frac{kT}{q} u_i + \frac{kT}{q} \xi_s F(u_i, \xi_s, u_F)$$

where $V_{FB}$ is the flat-band voltage, $C_i = \epsilon_i/\epsilon_s$ is the capacitance of the gate oxide layer of thickness $t_i$, $L_i = (\epsilon_i kT/2q^2 n_i)^{1/2}$ is the intrinsic Debye length, $u_F = (E_f - E_F)/kT$ is the normalized Fermi potential, and

$$F(u, \xi_s, u_F) = \left[ e^{u/(e^{-u} + u - 1)} + \frac{N_s}{n_i} \mathcal{F}_{3/2}(\xi_s) \right]^{-1/2}.$$  

(2)

Here, $N_s$ is the effective density of states in the semiconductor conduction band, $\mathcal{F}_{3/2}$ is the Fermi integral of the order 3/2 and $\xi_s = (E_F - E_{Fp})/kT$ is the normalized energy difference between the conduction band edge $E_F$ and the bulk Fermi level $E_F$.

In (2) and consequently in further considerations, Fermi–Dirac statistics are used as the semiconductor surface region in a transistor with a very thin gate oxide or with a high substrate doping concentration degenerates at a quite small excess value of the gate-source voltage above the threshold voltage. It has been proved [4] that errors resulting from the use of Maxwell–Boltzmann statistics in the analysis of such a transistor are small, but considerations in [4] have concerned only the drain current modeling accuracy.

The electron tunnel current density $I_{GC}$ for given values of $u_i$ and $\xi_e$ can be calculated using the model presented in Section II-B.

In order to perform the integration given in (1), the gradient $d\xi_e/du$ for any $\xi_e$ value must be known. It can be

found if the drain current is known. Provided $I_D >> I_{GC}$, which means that the lateral current $I(y)$ is practically constant along the channel, this current in any $y$ plane corresponding to a given $\xi_e$ value can be expressed in the form

$$I(y) = W \mu_{eff}(y) Q_n(y) \frac{kT}{q} \frac{d\xi_e}{du}$$

$$I_D = \frac{W kT}{L q} \int_{\xi_{FB}}^{\xi_e} \mu_{eff} \, d\xi_e$$

(3)

where $\mu_{eff}$ is the surface electron mobility and $Q_n$ is the surface charge density of electrons in the semiconductor in excess of the concentration at the depletion region edge $n(x_d) = n_i \exp \left(-u_F - \xi_s\right)$. $Q_n$ is calculated for given values of $u_i$ and $\xi_e$ by integrating the electron charge density over the semiconductor space-charge region

$$Q_n = -qN \cdot I_s$$

$$\int_0^{\xi_e} \frac{\mathcal{F}_{1/2}(\xi_s + u - \xi_e)}{F(u, \xi_s, u_F)} \, du.$$  

Then, extracting $d\xi_e/du$ from (3) and substituting it into (1) gives an expression for the gate electron tunnel current in the MOS transistor in the final form

$$I_{GC} = W L \int_{\xi_{FB}}^{\xi_e} \mu_{eff} Q_n J_{GC} \, d\xi_e$$

(4)

The hole tunnel current $I_{GH}$ is described by an analogical expression in which only $J_{GC}$ is replaced by $J_{GH}$. It is worth noticing that all three currents $I_{GC}, I_{GH}$, and $I_D$ can be calculated simultaneously in the same integration procedure.

B. Tunnel Current Densities

In the case of the gate–SiO₂–Si (100) system, the electron tunnel current density can be calculated using the one-dimensional integral form [5]

$$J_{GC} = q \int_{E_{min}}^{E_{max}} P(E_s) N(E_s) \, dE_s$$

(5)

where $E_s$ is the electron energy component associated with the motion direction normal to the barrier plane. The integral is over all energy levels within the semiconductor conduction band communicating with allowed states at the opposite side of the potential barrier, i.e., energy levels corresponding to the energy gap of the polysilicon gate are excluded.

$N(E_s)$ is the "supply function" and can be expressed by the following formula

$$N(E_s) = 4 \pi m^3 \ln(2) \frac{kT}{h^2} \left[ 1 + \exp \left( (E_{Fp} - E_s)/kT \right) \right]$$

- $\ln \left[ 1 + \exp \left( (E_{Fp} - E_s)/kT \right) \right]$
provided that changes in the electron effective mass across the MOS system are neglected [5] and the electron effective mass is represented by the scalar electron effective mass in the oxide \( m_e \). \( E_{F_m} \) is the Fermi level for electrons in the gate electrode, and \( E_F \) is the electron Fermi level at the semiconductor surface. Both of these magnitudes are counted in respect to the semiconductor conduction band bottom at the interface and are dependent on the surface potential \( \varphi_s \), and quasi-Fermi levels split \( \xi_n \).

The tunneling probability expressed by the exponential WKB factor

\[
P(E_i) = \exp \left[ -2 \int_{x_1}^{x_2} \kappa(x) \, dx \right]
\]

where \( x \) is the direction normal to the interface, \( x_1 \) and \( x_2 \) are classical turning points, and \( \kappa \) is the imaginary wave-number.

The hole tunnel current density \( J_{Gh} \) is modeled by an expression analogous to (5), but the integration is over energy states within the semiconductor valence band excluding energy levels corresponding to the polysilicon bandgap.

One of the questions in modeling tunnel currents in an MOS system is what dispersion relation to assume between the wave vector and the electron energy.

The most general Franz-type dispersion relation [6] with both oxide conduction band \( m_o \) and oxide valence band \( m_v \) effective masses can be written in the following form:

\[
\kappa = \left[ \frac{2m_o}{h^2} \frac{\phi(1 - \phi/E_o)}{1 - \left( 1 - \frac{m_o}{m_v} \right) \phi/E_o} \right]^{1/2}.
\]

Here, \( E_o \) is the oxide bandgap, and \( \phi \) is the difference between the oxide conduction band edge \( E_o \) (x) and the energy level \( E_x \) of the tunneling electron. Neglecting image forces, as is suggested in [7], and choosing \( x = 0 \) at the gate–oxide interface, we have

\[
\phi(x) = E_o(x) - E_x = \chi_x - qF_i(t_i - x) - E_x
\]

where \( F_i \) is the electric field in the oxide, \( \chi_x \) is the Si–SiO\(_2\) electron affinity, and as a consequence

\[
\int_{x_1}^{x_2} \kappa(x) \, dx = \left( \frac{2m_o}{h^2} \right)^{1/2} \frac{1}{qF_i} \left[ \frac{\phi(1 - \phi/E_o)}{1 - \left( 1 - \frac{m_o}{m_v} \right) \phi/E_o} \right]^{1/2} \phi \, dx
\]

where \( \phi_1 \) and \( \phi_2 \) are barrier heights at the classical turning points. The integration (8) can be performed analytically only at the assumption \( m_o = m_v = m_e \), as in the dispersion relation proposed in [8]. Then, taking advantage of the integration shown in [7], we have

\[
\int_{x_1}^{x_2} \kappa(x) \, dx = \left( \frac{2m_e}{h^2} \right)^{1/2} \frac{E_o^{1/2}}{16qF_i} \cdot \left( \sin \alpha_2 - \sin \alpha_1 - \alpha_2 + \alpha_1 \right)
\]

where \( \alpha_1,2 = 4 \arccos \left( \frac{\phi_1,2}{E_o} \right)^{1/2} \). In the case of Fowler–Nordheim tunneling, \( \phi_1 \) or \( \phi_2 = 0 \), and then \( \alpha_1 \) or \( \alpha_2 = 2\Pi \), respectively. Weinberg proves [7] that the parabolic dispersion relation, according to which

\[
\int_{x_1}^{x_2} \kappa(x) \, dx = \frac{2}{3} \left( \frac{2m_e}{h^2} \right)^{1/2} \frac{1}{qF_i} \left( \phi_2^{3/2} - \phi_1^{3/2} \right)
\]

provides the best description for the electron tunnel current in the Fowler–Nordheim regime \( \phi_1 \) or \( \phi_2 = 0 \) for any barrier height provided the single effective mass \( m_e = 0.5m_0 \) in SiO\(_2\) is assumed (where \( m_0 \) is the free electron mass). The parabolic dispersion relation does not take into account the existence of the valence band in the SiO\(_2\) (one-band barrier model), and this can be a source of underestimation in calculating results when the tunneling between energy states localized far from the oxide conduction band edge is considered. This is the case for tunneling between the gate and semiconductor valence band current \( J_{Gv} \).

Fig. 3 shows the ratio \( I_{Go}/I_{Gh} \) as a function of the gate-source voltage for an n-channel n+-polysilicon-gate MOS transistor at \( V_{DS} = 0 \). The curves are calculated using the model presented above with different approximations for the dispersion relation. It is assumed that \( \chi_x = 3.15 \) eV for the Si–SiO\(_2\) system and the Fermi level in the n+-polysilicon is located at its conduction band bottom. The results can be compared with experimental data of the ratio \( I_{Go}/I_{Gh} = (I_{Go} + I_{Gh})/I_{Gh} \) published in [9] and [10].

The shape of the curves can be easily interpreted. As the \( V_{GS} \) voltage increases and the substrate surface region is in the strong inversion state, the voltage across the oxide also strongly increases and the polysilicon conduction band bottom falls down relative to the substrate valence band top. When electrons in the substrate valence band "see" free states in the polysilicon conduction band, the current \( I_{Go} \) rapidly increases and the \( I_{Go}/I_{Gh} \) ratio decreases. Further increase of the gate voltage leads to the state (marked by full circles on the curves) in which the oxide conduction band bottom at the gate–oxide interface is in line with the electron Fermi level in the semiconductor substrate. Then Fowler–Nordheim tunneling of electrons from the substrate conduction band to the oxide conduction band dominates the electron tunnel current \( I_{Go} \). In this range of gate voltages, the ratio \( I_{Go}/I_{Gh} \) increases up to the state (marked by open circles) when the oxide conduction band bottom at the gate–oxide interface is in line with the substrate valence band top. Then Fowler–Nordheim tunneling of electrons from the substrate valence band to the oxide conduction band dominates the current \( I_{Go} \) and the ratio \( I_{Go}/I_{Gh} \) decreases.

The dotted curve in Fig. 3 is obtained assuming the parabolic relation, i.e., the one-band barrier model (10), with a single electron effective mass in the oxide \( m_e = 0.5m_0 \) [7]. Since this model underestimates the \( I_{Go} \) current, the minimal value of the ratio \( I_{Go}/I_{Gh} \) is much too high compared to the experimental data obtained for \( t_i = 3.5 \) nm in [9] and [10].

The dashed line in Fig. 3 describes the same case \( t_i =
Fig. 3. A theoretically calculated ratio $I_{Gc}/I_{Gv}$ versus the gate voltage assuming different models for the dispersion relation.

3.5 nm) using a two-band barrier model with a single effective mass (9) and assuming $m_F = 0.6365m_0$ and $E_{gi} = 8.64$ eV. These parameter values have been used in [11] for successfully fitting measurement results of Fowler-Nordheim electron tunnel current in the Al-SiO$_2$-Si system. Weinberg noticed [7] that, in case of the Fowler-Nordheim tunneling regime from Al into the SiO$_2$ conduction band, the integrals (9) and (10) give approximately the same values if $m_F = 0.6365m_0$, $E_{gi} = 8.64$ eV, $m_p = 0.5m_0$, $\phi_2 = 0$, and $\phi_1 = 3.15$ eV (the Al-SiO$_2$ barrier height), and therefore, both models giving (9) and (10) are equivalent for this case. However, as follows from a comparison of Fig. 3 (the dashed curve) with the results of [9] and [10], the two-band barrier model represented by (9) with the parameters used in [11] predicts in turn too a minimal value of the $I_{Gc}/I_{Gv}$ ratio.

The solid lines in Fig. 3 are obtained using the two-band barrier model given by (8) with $E_{gi} = 9.3$ eV [12], $m_{ci} = 0.602m_0$, and $m_{ci} = 0.772m_0$. The values of $m_{ci}$ and $m_{ci}$ has been chosen so that simultaneously:

1) the integral of (8) and the integral of (10) with $m_p = 0.5m_0$ give the same values if $d_1 = 0$ and $\phi_2 = 3.15$ eV are assumed, i.e., the model provides a good description for electron tunnel current in the Fowler-Nordheim tunneling regime, and

2) the minimal value of the $I_{Gc}/I_{Gv}$ ratio for $t_i = 3.5$ nm is in agreement with a minimal value of this ratio as measured in [9].

All further calculations are based on this approximation. It is interesting to notice that the curves shown by solid lines in Fig. 3 and all curves shown in subsequent figures will be almost unchanged if the two-band barrier model (see (9)) with a single effective mass $m_F = 0.611m_0$ and $E_{gi} = 10.51$ eV (ensuring fulfillment of both conditions mentioned above) is assumed in the calculations.

III. RESULTS OF THE THEORETICAL ANALYSIS

Fig. 4 shows currents $I_{Gc}$, $I_{Gv}$, and $I_D$ versus voltage $V_{GS}$ at $V_{DS} = 1$ V for an n-channel MOSFET with very thin gate oxide $t_i = 2.5$ nm, relatively high substrate doping concentration $N_A = 10^{17}$ cm$^{-3}$, and channel length $L = 1 \mu$m. The electron effective mobility $\mu_{eff}$ = 500 cm$^2$/V$\cdot$s constant along the channel is assumed for simplification of the calculations. As can be seen, the gate tunnel current in a wide voltage range (from a practical point of view) is many orders of magnitude lower than the drain current, even in the case of the extremely thin gate oxide.

Fig. 5 shows the oxide thickness dependence of the drain current and the electron tunnel current for two channel lengths. Although the gate tunnel current increases exponentially with decreasing oxide thickness, the ratio $I_{Gc}/I_D$ at $V_{GS} - V_T = 2$ V and $V_{DS} = 1$ V is still low, even for $t_i = 1.5$ nm. Such a thin oxide is not realized in a controllable way; therefore, it is a rather impractical case. When the channel length is reduced, the drain current reversely proportional to it increases while the gate tunnel current proportional to the gate area decreases. Therefore, the ratio $I_{Gc}/I_D$ decreases in proportion to the square of the channel length when it is reduced in accordance with the miniaturization trend.

Fig. 6 shows the gate tunnel current components and the drain current versus the drain-source voltage. The drain current increases with decreasing oxide thickness, causing a decrease of the electron tunnel current $I_{Gc}$ and of the electric field in the oxide. Due to the reduced voltage across the oxide near the drain region, the hole tunnel current $I_{Gh}$ also decreases.

In order to verify the model quantitatively, it was used
MAKUSIAK: GATE TUNNEL CURRENT IN MOS TRANSISTORS

for the generation of drain current and gate tunnel current versus the gate-source voltage characteristics for the n⁺-polysilicon-gate n-channel MOS transistor investigated experimentally in [2]. A comparison between theoretical and experimental results is shown in Fig. 7. The dependence of the electron effective mobility on the effective transverse electrical field in the channel was modeled using the approximation given in [13]. A flat-band voltage was used as a fitting parameter. The experimental subthreshold drain current slope is lower than the value predicted theoretically, which can be a result of short-channel effects ($L = 0.15 \mu m$) not included in the theoretical model. An experimental gate tunnel current level was obtained in theoretical calculations after assuming the electron affinity $\chi_e = 2.9$ eV for the Si-ultrathin SiO₂ system ($t_i = 2.5$ eV). This value is a little lower than that usually accepted for a Si-SiO₂ system, but it is reasonable taking into consideration an experimentally observed tendency according to which $\chi_e$ decreases in regard to its thick-oxide value when the oxide thickness falls below about 3.5 nm [14], [15]. On the other hand, this value agrees with the barrier height used in [7] after the energy of the first subband is subtracted from the full barrier of 3.1 eV. The slope of the theoretical gate tunnel current curve is a little smaller in comparison with the experimental curve. However, considering the fact that the electron-physical parameters of the MOS system with ultrathin oxide are not well known and that other effects, which can result from, e.g., very high substrate doping concentration $N_A = 1.5 \cdot 10^{18}$ cm$^{-3}$ and very short channel $L = 0.15 \mu m$, are not considered in the model, the agreement seems to be quite good.

IV. CONCLUSIONS

We show in this paper how to calculate the gate tunnel current in an MOS transistor. Details of the model such as the description of the electron charge $Q_e$ and the densities of the tunnel currents $J_{G,t}$, can be changed as a result of a compromise between simplicity and accuracy. Here, because of the need to present a theoretical analysis in the work, the integral formulas for these magnitudes have been used.

A comparison between the gate tunnel current and the drain current based on the theoretical model shows that the gate tunnel current is low in comparison with the drain current, even in the case of extremely thin gate oxides. This statement is in agreement with experimental results published in the literature. In conclusion, one can suppose, therefore, that the minimal gate oxide thickness in MOS/VLSI circuits will result from the ability to realize very good quality ultrathin oxides in a controllable way rather than from the flow of too large a gate tunnel current. However, in some circuit applications of the MOS transistor the problem of input leakage current can be especially important, and then the gate tunnel current can limit the minimal oxide thickness.
REFERENCES


Bogdan Majkusiak was born in Warsaw, Poland, on February 16, 1955. He received the M.Sc. and Ph.D. degrees from the Department of Electronics, Technical University of Warsaw, in 1979 and 1985, respectively.

Since 1978, he has been working for the Institute of Electron Technology (now the Institute of Microelectronics and Optoelectronics), Technical University of Warsaw. His research activity has concentrated in the area of the theory and technology of metal–insulator–semiconductor devices, especially with very thin and ultrathin oxides.