Unified compact theory of tunneling gate current in metal–oxide–semiconductor structures: Quantum and image force barrier lowering

Wu-yun Quan and Dae M. Kim[a)]
Korea Institute for Advanced Study, 207-43 Cheongryangri-dong, Dongdaemun-gu, Seoul 130-012, Korea

Myoung Kwan Cho
Memory Division, Semiconductor Business, Samsung Electronics, Co., Ltd., Kyungki-do 449-900, Korea

(Received 22 April 2002; accepted for publication 9 July 2002)

A compact model of gate current due to Fowler–Nordheim tunneling is presented, which agrees closely with the self-consistent numerical analyses of the surface inversion region of metal–oxide–semiconductor field-effect transistors (MOSFETs). The model can quantify the measured data with the accuracy practically identical to the time consuming numerical simulation. It is also shown conclusively that image force lowering of the oxide barrier height is negligible for the oxide as thin as 1 nm. The quantum barrier lowering resulting from subband splitting is rigorously incorporated, including the effect of two-dimensional electrons inverted at the higher lying subbands. Finally, it is pointed out that the compact model can be readily generalized to include the direct tunneling in deep submicron MOSFETs. © *2002 American Institute of Physics.* [DOI: 10.1063/1.1504173]

I. INTRODUCTION

The increasing gate current (I_G) with the use of ultrathin gate oxides is a key issue limiting metal–oxide–semiconductor field-effect transistor (MOSFET) scaling and it is important to understand and specify I_G in terms of the device and material parameters. A crucial parameter affecting I_G is the effective barrier height (Φ_{eff}) of the oxide, which is encountered by the tunneling electrons. Φ_{eff} is generally determined by two factors, *viz.* the image force lowering,[1–12] and the quantum lowering arising from the quantized energy subbands in the surface potential well.[12–18] The latter effect was originally quantified in the triangular potential well (TPW) approximation[13] but numerical simulations based on the self-consistent analyses of the coupled Poisson and Schrödinger equations have made it possible to characterize the quantum effects accurately.[14] However, the image-force barrier lowering in MOS structures still remains a hazy concept. This is evidenced by a number of papers in the literature, which incorporate the image force for analyzing I_G,[1–12] while many other articles neglect the effect.[13–18]

The purpose of this article is to present an analytic compact model of gate current due to Fowler–Nordheim (F–N) tunneling. The model can quantify the measured data with the accuracy practically identical to the time consuming numerical simulation. Also the article examines comprehensively the image-force lowering as applied to MOS structures. The quantum barrier lowering is considered in detail by using the numerical simulation including the effect of two-dimensional (2D) electrons inverted in the higher lying subbands. The effective barrier height thus clarified is used for formulating a unified compact gate current model in deep submicron MOSFETs.

a)Electronic mail: dmkim@kias.re.kr

II. COMPACT GATE CURRENT MODEL DUE TO F–N TUNNELING

A. Existing F–N tunneling models

The most widely used model of F–N tunneling gate current in the MOS structure is given in terms of the effective barrier height (Φ_{eff}) and the surface electric field (F_S) as[2,15]

$$J_G = J_0 \exp\left(-\frac{\theta}{F_S} \Phi_{\text{eff}}^{3/2} \right), \tag{1}$$

where $\theta = 4\epsilon_{\text{OX}}\sqrt{2m_{\text{OX}}}/3q\hbar\epsilon_{\text{Si}}$ is constant given in terms of the electron effective mass in oxide, m_{OX}, the permittivities of silicon and oxide, ϵ_{Si}, ϵ_{OX}, the electron charge, q, and the Planck constant, $\hbar$. The pre-exponential factor (J_0) is proportional to F_S^2:

$$J_0 = C_0 F_S^2 = \frac{q^3 \epsilon_{\text{Si}}^2}{16\pi^2 \hbar \epsilon_{\text{OX}}^2} \left(\frac{m_0}{m_{\text{OX}}} \right) \frac{F_S^2}{\Phi_{\text{eff}}}. \tag{2}$$

Various values of electron effective mass (m_{OX}) in the oxide have been reported[2,3,15] but we use in this article the value of $0.5m_0$, where m_0 is the free-electron mass, following the discussion by Weinberg.[15]

As clear from Eq. (1) the gate current is critically dependent on the effective barrier height. Φ_{eff} is often treated as constant, and the typical value of Φ_{eff} is given by Weinberg as 2.9 eV.[15] However, Φ_{eff} is generally considered to be determined by two factors, *viz.* the image-force lowering, and the quantum lowering due to the quantized energy subbands in the surface potential well. The former is frequently quantified by $\Delta\Phi_{\text{IMG}} = \alpha F_S^{1/2}$ with α a constant parameter.[4–6] The quantum lowering of the barrier would obviously be different for electrons residing in different subbands but the barrier is usually treated with one-subband approximation in the TPW analysis as $\Delta\Phi_{\text{QM}} = \beta F_S^{2/3}$, where β is a constant.[7,13,15] In summary, Φ_{eff} is represented in the literature in the form

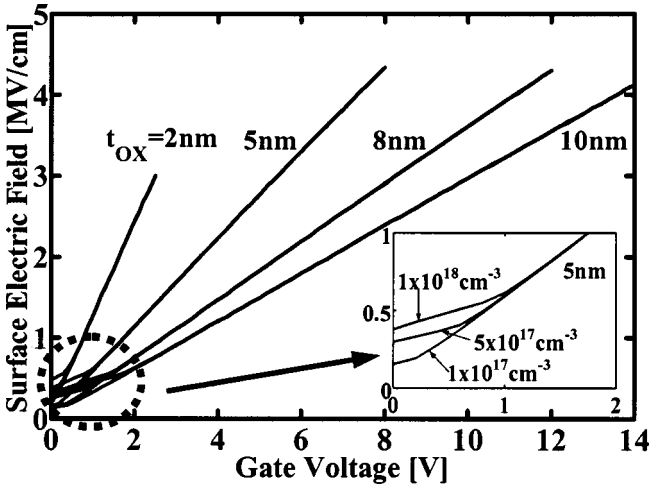

FIG. 1. Surface electric field vs gate voltage for four different oxide thicknesses of 2, 5, 8, and 10 nm. The calculations were done on ⟨100⟩ silicon substrate by self-consistently solving the coupled Poisson and Schrödinger equations for given oxide thickness. Three *p*-substrate doping levels of $1\times10^{17}\,\mathrm{cm}^{-3}$, $5\times10^{17}\,\mathrm{cm}^{-3}$, and $1\times10^{18}\,\mathrm{cm}^{-3}$ were considered with n^+-poly gate doping concentration of $1\times10^{21}\,\mathrm{cm}^{-3}$ in all cases. The surface electric field is only affected slightly in the depletion region (small gate voltage) as indicated in the insert for 5 nm gate oxide.

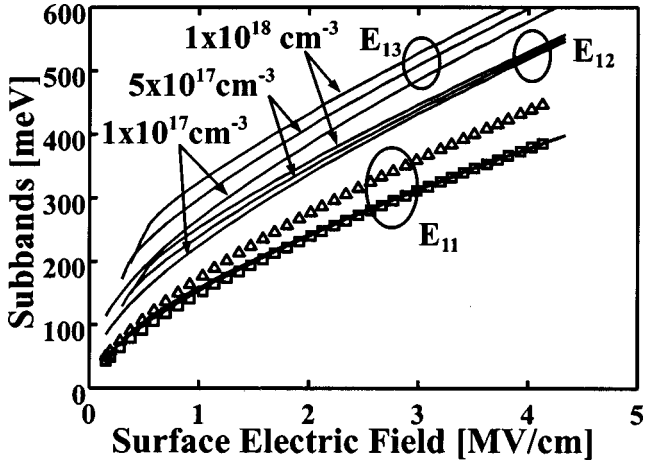

FIG. 2. Subbands vs surface electric field for three substrate-doping concentrations used in Fig. 1 (solid lines). Shown also is the ground subband calculated by triangular potential well approximation, *viz.* $E_{11}=\beta F_S^{2/3}$, with $\beta=0.17$ [eV/(MV/cm)$^{2/3}$] (open triangles). The numerically simulated E_{11} is shown independent of substrate doping and can be accurately fitted by $E_{11}=\beta F_S^{2/3}$, but with $\beta=0.15$ [eV/(MV/cm)$^{2/3}$] (open squares).

$$\Phi_{\mathrm{eff}}=\Phi_B-\alpha F_S^{1/2}-\beta F_S^{2/3}, \tag{3}$$

where $\Phi_B=3.15\,\mathrm{eV}$[3,14] is the conduction band discontinuity at the Si–SiO_2 interface. The typical coefficients reported were $\alpha=0.259$ [eV(MV/cm)$^{-1/2}$][4–6] and $\beta=0.1$ [eV(MV/cm)$^{-2/3}$].[4]

The models discussed earlier predicted the F–N tunneling current independent of the substrate doping, which will be proven correct by experimental data and theoretical analyses. However, these models could not satisfactorily quantify the measured gate current, as will be detailed later.

B. Compact model of F–N tunneling current

We present a compact model, capable of accurately quantifying the measured gate current due to F–N tunneling in MOS structures. The model is based on judicious use of the self-consistent numerical solutions of the coupled Poisson and Schrödinger equations[16,17] whose results have two basic features.

The first feature concerns with the effect of the substrate doping. Figure 1 shows the surface electric field versus the applied gate voltage as obtained by the self-consistent numerical simulation. In the analysis four samples were considered with oxide thickness of 2, 5, 8, and 10 nm, respectively. For each oxide, three doping levels of $1\times10^{17}\,\mathrm{cm}^{-3}$, $5\times10^{17}\,\mathrm{cm}^{-3}$, and $1\times10^{18}\,\mathrm{cm}^{-3}$ for the *p*-type substrate were used, and the doping concentration of $1\times10^{21}\,\mathrm{cm}^{-3}$ for the n^+-poly gate was used in all cases considered. The calculations were performed for ⟨100⟩ silicon substrate. Clearly, the surface electric field becomes independent of the substrate doping in strong inversion, as anticipated, since the bulk capacitance is in this case mainly determined by the inversion capacitance. This enables the F–N tunneling current to be expressed as a function of the surface electric field, independent of substrate doping.

The second feature concerns with the quantum barrier lowering due to the subband splitting. Figure 2 shows the three subbands, E_{11}, E_{12}, and E_{13}, versus the surface electric field, F_S, which were found from the self-consistent numerical analysis. Also shown for comparison is the ground subband found from the TPW approximation.[7] As anticipated, the TPW result overestimates E_{11}, since the triangular potential well has narrower width than the numerically calculated potential well. More importantly, the self-consistent results show that only the ground subband is practically independent of the substrate doping, and is to be fitted with excellent degree of approximation by the 2/3 power law of F_S, *viz.* $E_{11}=\beta F_S^{2/3}$, with $\beta=0.15$ [eV/(MV/cm)$^{2/3}$]. This power law is fortuitously the same as the analytical result of the TPW approximation but in the latter theory β is larger by about 13%, i.e., $\beta=0.17$ [eV/(MV/cm)$^{2/3}$].[7] Although the TPW approach predicts that all subbands follow the 2/3 power law, and are independent of substrate doping, the self-consistent numerical results show that the higher subbands do not follow this power law, and depend on the substrate doping.

To summarize, the gate current density, J_G, is quantified in the present model in terms of F_S and the F–N tunneling at the ground subband, which renders J_G independent of the substrate doping, in accordance with the experimentally observed tendency. The modification arising from the tunneling at the higher subbands is rigorously incorporated in the preexponential factor in Eq. (1), as discussed later.

The gate current density due to F–N tunneling of the 2D electrons is generally given by

$$J_G=\sum_{i,j} qN_{ij}f_{ij}P_{ij}, \tag{4}$$

where *i* and *j* denote the valley index and the quantum number respectively, N_{ij}, f_{ij}, and P_{ij} are the 2D electron density, the frequency with which an electron strikes the inter-

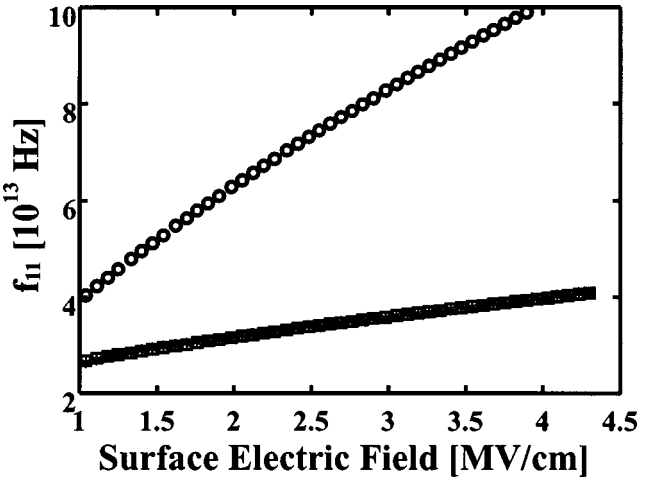

FIG. 3. Frequency (f_{11}) with which an electron in ground subband strikes the interface vs surface electric field. The stars denote the numerically calculated values, whose fitting by Eq. (8) is also shown (open squares). The result of the triangular potential well approximation (open circles) is also shown for comparison.

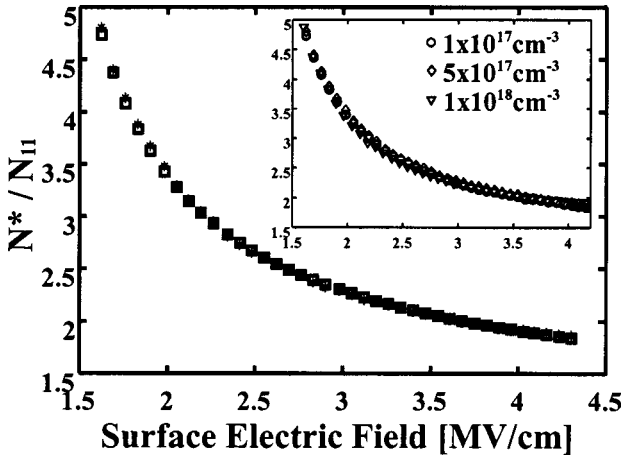

FIG. 4. The ratio of the effective 2D electrons inverted in all subbands with respect to the ground subband electron density, N^*/N_{11} vs surface electric field. The stars denote the numerically calculated results, whose fitting with Eq. (9) is also shown (open squares). Shown in the inset is the numerically calculated ratio for different doping concentrations in Fig. 1.

face, and the tunneling probability of electrons in the corresponding subbands (E_{ij}), respectively. Eq. (4) is rewritten as

$$J_G = qN^* f_{11} P_{11} \tag{5}$$

with

$$N^* = \left(1 + \sum_{ij \neq 11} \frac{N_{ij} f_{ij} P_{ij}}{N_{11} f_{11} P_{11}}\right) N_{11}. \tag{6}$$

The 2D electrons inverted in the higher subband, N_{ij}, are scaled with respect to N_{11} with P_{ij}, f_{ij} used as weighting factors. In this way the F–N tunneling of the entire inverted electrons in all subbands is rigorously accounted for in terms of P_{11} and f_{11}.

Now, f_{11} as a function of F_S can be best investigated by considering triangular potential well approximation as an example. By definition one can write[7,18]

$$f_{11} = \left(2 \int_0^{z_{\text{turn}}} \frac{dz}{\nu_z}\right)^{-1}, \tag{7}$$

where ν_z is the electron velocity in the z direction normal to the interface in the triangular potential, i.e., $m_{11}\nu_z^2/2 + qF_S z = \beta F_S^{2/3}$ and $z_{\text{turn}} = \beta / qF_S^{1/3}$ the classical turning point. The integration of Eq. (7) readily predicts $f_{11} \propto F_S^{2/3}$. In Fig. 3 f_{11} thus found is plotted as a function of F_S, together with the exact result found numerically by solving the coupled Poisson and Schrödinger equations. The TPW result departs appreciably from the exact numerical result due obviously to the overestimation of E_{11} and inaccuracy of the potential width. However, as shown in Fig. 3 the accurate fit to the exact numerical result can be made with a simple analytic expression, which is still dictated by the 2/3 power law in the form

$$f_{11} = (0.87 F_S^{2/3} + 1.8) \times 10^{13} \ \text{Hz} \tag{8}$$

with F_S expressed in MV/cm. Note that Eq. (8) can generally be used and transcribed into the gate voltage, irrespective of the substrate doping and the oxide thickness (see Figs. 1 and 2).

The effective 2D inversion electrons, N^* was also found exactly by numerical means and is plotted versus F_S in Fig. 4. The ratio, N^*/N_{11} becomes as large as about 5 for small V_G or F_S because of (a) the fraction of electrons inverted in the higher lying subbands is relatively large in this voltage regime and (b) the tunneling probability at higher subbands is large compared with P_{11}. With increasing V_G, although most of the inversion electrons reside in the ground subband, the small fraction of electrons lying in the higher subbands still contributes to the gate current substantially because of their high tunneling probabilities. The analytical function capable of fitting accurately the numerically obtained ratio, N^*/N_{11}, is given by

$$\frac{N^*}{N_{11}} = 1 + \frac{20 F_S^{-4} + 7}{2.2 F_S - 0.9}, \tag{9}$$

where F_S expressed in [MV/cm]. Equation (9) is also plotted and compared with the exact numerical result in Fig. 4. Also shown in the figure are the numerical analyses indicating that the dependence of this ratio on the substrate doping is very weak in the range from $1 \times 10^{17}\,\text{cm}^{-3}$ to $1 \times 10^{18}\,\text{cm}^{-3}$, and Eq. (9) can also be generally used just like Eq. (8). The specification of N^* is thus complete with the inversion 2D electron in the ground subband given by[7]

$$N_{11} = \frac{2 g_1 m_{d1} k_B T}{\pi \hbar^2} \ln[1 + e^{-(E_{11} - E_{FP})/k_B T}], \tag{10}$$

where g_1, m_{d1} are the degeneracy and the effective electron density-of-state mass of the lower valley, respectively, E_{FP} the substrate Fermi energy level.

By combining Eqs. (5), (8), (9), and (10) the gate current density reads as

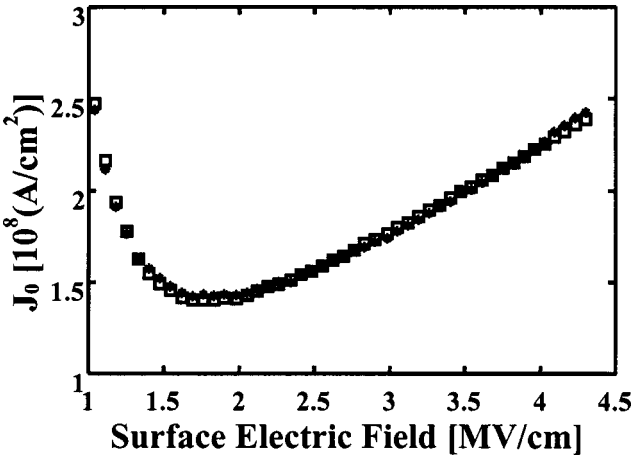

FIG. 5. The pre-exponential factor in Eq. (12) vs surface electric field (open squares). Also shown are the numerically simulated results for comparison (stars).

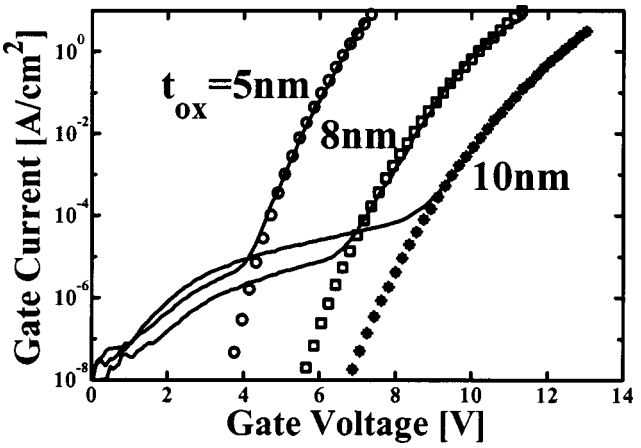

FIG. 6. The gate current vs gate voltage (solid lines) measured from n^+-poly gate NMOSFETs with oxide thickness of 5, 8, and 10 nm, respectively. The symbols represent the theoretical fits to the F–N tunneling component of the data by using the model developed in this work.

$$J_G = J_0 \exp\left(-\frac{\theta}{F_S} \Phi_{\text{eff}}^{3/2} \right) \tag{11}$$

with the pre-exponential factor given in terms of F_S as

$$J_0 = A(0.87F_S^{2/3} + 1.8)\left(1 + \frac{20F_S^{-4} + 7}{2.2F_S - 0.9} \right) k_B T \times \ln[1 + e^{-(E_{11} - E_{FP})/k_B T}], \tag{12}$$

where $A = 2qg_1 m_{d1}/\pi\hbar^2$, and F_S is in MV/cm. The compact model of Eq. (11) specifies the gate current due to F–N tunneling of all electrons inverted in ground and higher subbands. The pre-exponential factor of Eq. (12) is plotted versus F_S in Fig. 5 and is compared with the exact numerical result. As clear from the figure, the agreement between the two is excellent. For small V_G or F_S, J_0 becomes large, since a substantial fraction of electrons is inverted in higher subbands having higher tunneling probabilities. With V_G increasing the fraction of electrons in higher subbands rapidly decreases, reducing thereby J_0. With further increase of V_G nearly all of the inverted electrons distribute in the ground subband but the number of electrons therein increases with V_G and J_0 should increase.

In the next section the effective barrier potential operative at the ground subband is systematically examined in the light of the measured data.

III. COMPARISON WITH MEASURED DATA AND PARAMETER EXTRACTION

The new compact model of the F–N tunneling gate current [Eq. (11)] is next compared with the measured gate current. The devices used for the measurement were n^+-poly gate NMOSFET devices fabricated on ⟨100⟩ silicon substrate. Three samples were used with respective oxide thickness of 5, 8, and 10 nm and corresponding substrate doping concentrations of about $5\times10^{17}\,\text{cm}^{-3}$, $3.5\times10^{17}\,\text{cm}^{-3}$, and $1\times10^{17}\,\text{cm}^{-3}$, respectively.

Figure 6 shows the gate current versus V_G measured from the three samples. These data are used to extract and/or test the parameters α and β determining the effective barrier potential, Φ_{eff} in Eq. (1). Φ_{eff} as represented by Eq. (3) is a function of F_S and F_S in turn depends on V_G and oxide thickness (t_{OX}) (see Fig. 1). Hence, J_G is in effect a function of V_G, t_{OX}, with α and β for Φ_{eff} constituting parameters, i.e., $J_G = J(V_G, t_{\text{OX}}, \alpha, \beta)$.

Although α and β in Eq. (3) are to be viewed rooted in the image force and quantum lowering of the barrier height, these effects have to be tested self-consistently via the best fit to the data. The fitting is to be routinely done, using the method of minimum sum of the mean square deviation for given t_{OX}. More ambitiously, however, t_{OX} can also be regarded as a parameter to be determined from the best fit and the resulting electrical thickness of t_{OX} can be compared with transmission electron microscopy observation for consistency. The latter approach was chosen for the data fitting and the results are shown in Fig. 6, and the optimal parameters t_{OX}, α, β thus extracted are given in Table I. Obviously the compact model of Eq. (11) can fit the data with high degree of accuracy. In analyzing the data the F–N component of the gate current was considered only and the regions of stress induced leakage current or direct tunneling current were omitted in our analysis.

In Fig. 7 the same gate currents in Fig. 6 are plotted as a function of the surface electric field, and the open squares are the result of the present compact model [Eq. (11)] with the use of α, β optimally determined, i.e., $\alpha=0$ and $\beta=0.15$ $[\text{eV}/(\text{MV/cm})^{2/3}]$. In this case the three $I_G - V_G$ curves corresponding to three different data from samples with different oxide thickness are shown to collapse into the single univer-

TABLE I. The extracted parameters α, β, and the oxide thickness (t_{OX}) from the three sample devices.

t_{OX} (Å)		α	β $[\text{eV}/(\text{MV/cm})^{2/3}]$
Design	Extracted		
50	52.5	0	0.156
80	83	0	0.150
100	103	0	0.162

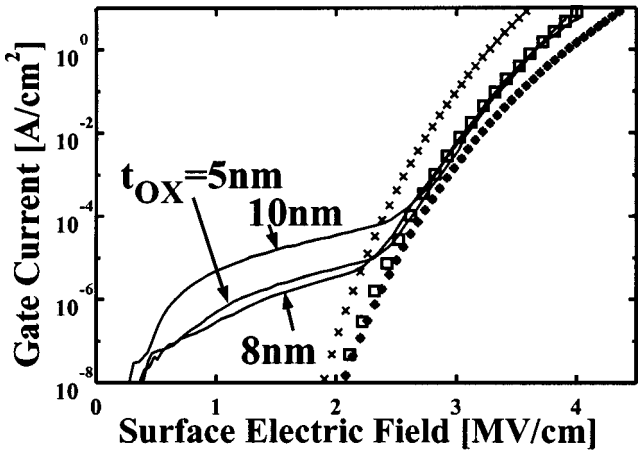

FIG. 7. The gate current vs surface electric field. The solid lines represent the data in Fig. 6, the open squares the compact model given by Eq. (11). Also shown for comparisons are the existing models for $\Phi_{\text{eff}}=2.9$ eV (see Ref. 15) (stars), and for Φ_{eff} given by Eq. (3) with $\alpha=0.259$ [eV(MV/cm)$^{-1/2}$] (see Refs. 4–6) and $\beta=0.1$ [eV(MV/cm)$^{-2/3}$] (see Ref. 4) (crosses).

sal I_G-F_S curve, as it should. Also shown in Fig. 7 are the results of the previous classical analytical models for the cases of constant Φ_{eff}[3,14] and Φ_{eff} varying according to Eq. (3) with reported α, β.[4] As clear from this figure there is a substantial disagreement between the existing models and the measured data. In the next section the lowering of Φ_{eff} is discussed in detail.

IV. IMAGE FORCE AND QUANTUM LOWERING OF BARRIER HEIGHT

As shown in Table I, the optimal value of α for best fitting the data in Fig. 7 is zero, which implies that there is no image-force lowering in MOS structures. This finding is in contrast with the reported value of $\alpha=0.259$ [eV(MV/cm)$^{-1/2}$][4–6] for MOSFET structures and should be carefully examined.

The image-force lowering was first introduced in conjunction with Schottky contact in semiconductor devices, as illustrated in Fig. 8(a).[19] Note from this figure that in order for the image-force lowering to be appreciable three conditions should be satisfied: (i) the metal-like image plane, (ii) the barrier and image potentials being on the same side, and (iii) the barrier height located near the image plane. When these conditions are met as in the case of Schottky contact, the effect of image force lowering is significant.

In MOS structures the first two conditions are satisfied with the poly-gate/oxide interface constituting the metal-like image plane. However, unlike the case in Schottky contact, the barrier height is removed from the image plane by oxide thickness.

Now the image potential, $\phi_I(z)$, in uniform dielectric media with the relative permittivity ϵ_r, is given by[19]

$$\phi_I(z)=-\frac{q^2}{16\pi\epsilon_0\epsilon_r z}, \tag{13}$$

where z is the distance away from the image plane, and ϵ_0 is the vacuum permittivity. Since ϵ_r in the oxide is smaller than

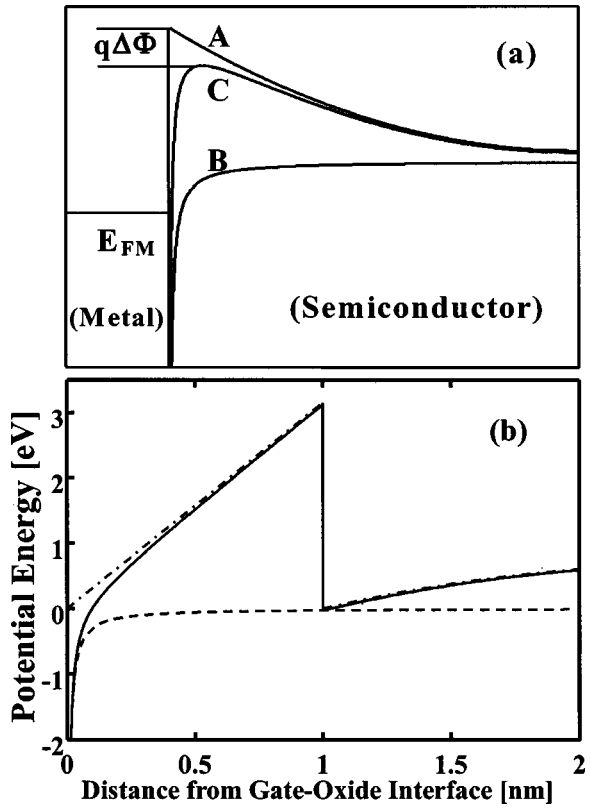

FIG. 8. (a) Schematics for the image force barrier lowering in metal/semiconductor contact: the metal/semiconductor contact potential energy (A), the image potential energy (B), and the total potential energy (C). E_{FM} denotes the metal Fermi energy level. (b) Barrier height lowering due to the image force in MOS structure: the dashed line is the image potential; the dot-dashed line the potential in oxide and substrate; the solid line the total barrier potential.

that in silicon, one can estimate the maximum possible image force lowering by putting $\epsilon_r=\epsilon_{\text{OX}}$. The image potential of Eq. (13) is plotted in Fig. 8(b) and it is clear from this figure that for MOS structures, the image force lowering becomes negligible even for the oxide thickness as thin as 1 nm. This is in complete agreement with our result, i.e., $\alpha=0$.

Additionally, the same value of β {$=0.15$ [eV/(MV/cm)$^{2/3}$]} can fit both the measured gate current and the ground subband found exactly by numerical means. This clearly suggests that the barrier lowering in MOS structures is solely due to splitting of subbands in the surface charge region [see Eq. (3)].

V. CONCLUSIONS

We have presented in this article a compact, analytical model of the gate current due to F–N tunneling in deep submicron MOSFETs. The model accounts for the entire electrons inverted in all subbands and participating in the tunneling and can quantify the measured data with the accuracy identical to the time-consuming numerical simulation. We have also shown conclusively that the image force lowering of the oxide barrier height is negligible for the gate oxide as thin as 1 nm and the effective barrier is, therefore, determined solely by the quantum effect. Although the model

is developed for the gate current due to F–N tunneling, the basic results can readily be extended to describe the gate current due to direct tunneling or the combination of F–N and direct tunneling by simply adjusting the tunneling probabilities.

[1] R. E. Burgess, H. Kroemer, and J. M. Houston, Phys. Rev. **90**, 515 (1953).
[2] M. Lenzlinger and E. H. Snow, J. Appl. Phys. **40**, 278 (1969).
[3] A. Schenk and G. Heiser, J. Appl. Phys. **81**, 7900 (1997).
[4] T. H. Ning, C. M. Osburn, and H. N. Yu, J. Appl. Phys. **48**, 286 (1977).
[5] S. Tam, P. K. Ko, and C. M. Hu, IEEE Trans. Electron Devices **31**, 1116 (1984).
[6] C. M. Yih, S. M. Cheng, and S. S. Chung, IEEE Electron Device Lett. **45**, 2343 (1998).
[7] T. Ando, A. Fowler, and F. Stern, Rev. Mod. Phys. **54**, 437 (1982).
[8] J. Schewchun and V. A. K. Temple, J. Appl. Phys. **43**, 5051 (1972).
[9] D. R. Young, J. Appl. Phys. **47**, 2098 (1976).
[10] T. H. Ning, Solid-State Electron. **21**, 273 (1978).
[11] M. V. Fischetti and S. E. Laux, J. Appl. Phys. **78**, 1058 (1995).
[12] C. M. Yih, S. M. Cheng, and S. S. Chung, IEEE Electron Device Lett. **45**, 2343 (1998).
[13] Z. A. Weinberg, Solid-State Electron. **20**, 11 (1977).
[14] F. Rana, S. Tiwari, and D. A. Buchanan, Appl. Phys. Lett. **69**, 1104 (1996).
[15] Z. A. Weinberg, J. Appl. Phys. **53**, 5052 (1982).
[16] S. H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, IEEE Trans. Electron Devices **18**, 209 (1997).
[17] S. H. Lo, D. A. Buchanan, and Y. Taur, IBM J. Res. Dev. **43**, 327 (1999).
[18] L. F. Register, E. Rosenbaum, and K. Yang, Appl. Phys. Lett. **74**, 457 (1999).
[19] S. Sze, *Physics of Semiconductor Devices*, 2nd ed. (Wiley, New York, 1981), p. 250.