Simulation of a quantum-dot flash memory

Giuseppe Iannaccone
Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni, Università di Pisa
A. Trellakis
Beckman Institute and Department of Electrical and Computer Engineering, University of Illinois
U. Ravaioli
Beckman Institute and Department of Electrical and Computer Engineering, University of Illinois

G. Iannaccone, A. Trellakis, U. Ravaioli, *Simulation of a quantum-dot flash memory*, Journal of Applied Physics, **84**, pp.5032-5036 (1998).

Simulation of a quantum-dot flash memory

G. Iannaccone
Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica e Telecomunicazioni, Università degli studi di Pisa, Via Diotisalvi 2, I-56126 Pisa, Italy

A. Trellakis and U. Ravaioli[a)]
Beckman Institute and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801

(Received 9 April 1998; accepted for publication 4 August 1998)

We present the simulation of a flash memory in which the floating gate is replaced by a silicon quantum dot. Unlike conventional flash memories, this device promises the advantage of self-limited direct charging of the floating gate with a low writing voltage, thus allowing one to overcome hot carrier degradation problems. Due to the small dimensions of this memory device, quantum effects are expected to play an important role. To estimate their magnitude, we compare a semiclassical simulation based on the solution of Poisson's equation with a quantum computation solving the complete system of Schrödinger's and Poisson's equation. Despite the three-dimensional nature of the problem, a two-dimensional computational mesh proved to be sufficient to provide good agreement with experimental results, after three-dimensional corrections were included.
© *1998 American Institute of Physics*. [S0021-8979(98)06221-5]

I. INTRODUCTION

The long-term future of semiconductor electronics depends very much on the development of nanoscale devices. However, the huge capital investment into current process technology discourages radical changes. Therefore, an evolutionary approach to nanoelectronics may very well be the safest and most cost-efficient approach. A first step in such an approach would consist of addressing nanoscale effects as they emerge in the process of scaling down conventional device geometries; a second step in trying to exploit these effects in order to improve device functionality and performance, while changing conventional materials technology as little as possible.

Recent experiments indicate that it is possible to fabricate flash memories with one or several quantum dots in the place of the floating gate, taking advantage of Coulomb blockade effects emerging even at room temperature, as capacitances are scaled down with device dimensions.

In 1996, the first nanocrystal based memory was demonstrated,[1,2] requiring only the addition of few process steps to conventional complementary metal–oxide–semiconductor (CMOS) technology. In that case, the floating gate was replaced by a distributed film of nanocrystals of silicon with an average diameter of 5 nm, and a sheet density of $(0.3-1)\times 10^{12}$ cm^{-2}. More recently there have been reports[3–5] of flash memories fabricated with a single quantum dot in the place of the floating gate, realized with silicon on insulator (SOI) technology. For those devices, the floating gate is either a polysilicon dot with estimated dimensions 7 nm×7 nm×2 nm,[3] or a polysilicon dot with a size of about 30 nm×30 nm×25 nm,[4] or an amorphous silicon dot of 30 nm×30 nm×8 nm.[5]

With respect to current flash memories, these devices have the important advantage of not requiring hot carriers for charging the floating gate. In fact, electrons tunnel directly from the channel to the floating gate when an adequate positive voltage is applied on the control gate. Also, given the small capacitance between the control gate and the dot, only a few electrons are needed for shifting the threshold voltage significantly. In this way, hot carrier degradation problems are removed and, in particular for the device of Ref. 5, low writing voltages seem to be sufficient. Furthermore, the charging process is self-limited because of Coulomb repulsion.

On the other hand, the thinner oxide required for direct tunneling from the channel to the dot poses leakage problems and may significantly decrease the refresh time. The tradeoff between writing/erasing times and the refresh time is a major issue for fully integrating these devices with current silicon memory technology.

In this paper we perform a numerical simulation of a quantum-dot flash memory in order to extract the relevant parameters for device operation and to understand the relationship between details of the structure and the meaningful quantities, as required for an efficient design from the electrical point of view.

We focus on a device similar to that fabricated by Welser *et al.*,[5] which exhibits a shift in the threshold voltage of about 0.75 V at room temperature and a low write/erase voltage of ± 3 V. The layer structure is shown in Fig. 1 where, starting from the bottom, we have the following layers: a 25 nm layer of silicon on insulator substrate, 2.5 nm tunnel oxide, 8 nm of amorphous silicon, and 9 nm of control oxide. The 50 nm wide silicon channel is patterned by electron beam lithography and successive etching, then the poly-

[a)]Author to whom correspondence should be addressed; electronic mail: ravaioli@uiuc.edu

 © 1998 American Institute of Physics

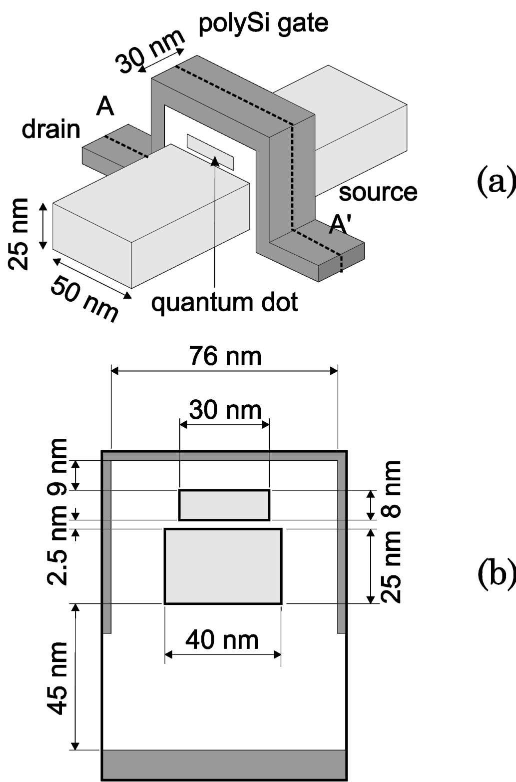

FIG. 1. The device structure (a) and its cross section A-A' (b). Dark gray regions are polysilicon, light gray regions are silicon, white regions are SiO_2.

silicon gate is deposited and patterned in the direction perpendicular to the channel.

The silicon dot and the channel have been oxidized during the growth of the lateral oxide, before polysilicon deposition. Thus, based on transmission electron microscope (TEM) images of the structure shown in Ref. 5, we assume that the actual dimension of the self-aligned dot is 30×30 nm^2 and that the Si channel is constricted to 40 nm under the gate [Fig. 1(b)].

II. SIMULATION

Simulation of this class of devices is computationally quite demanding, because the device is three dimensional and the effect of the limited width of the double-gate MOS cannot be discarded. In our approach, we attempted a simplification of the problem by using a two-dimensional computational model together with three-dimensional corrections outlined below, which appear to give a satisfactory description of experimental results. Other issues arise from the small device dimensions which necessitate the careful evaluation of quantum confinement effects both in the dot and in the inversion layer.

In order to deal with this latter issue, we have performed both a semiclassical and a quantum-mechanical simulation. In the semiclassical simulation electrons in the dot and in the channel obey a distribution simply determined by the density of states in the bulk and by Fermi–Dirac statistics; in the quantum-mechanical case, the allowable states in the dot and in the channel are obtained by solving the Schrödinger equation on the transverse section, where the density of states has the typical behavior of one-dimensional structures. As we shall show, the density of states affects the value of the threshold voltages and other electrical device parameters. The coupled Poisson–Schrödinger equations are solved using the recently developed iteration scheme described in Ref. 6.

There are two effects that must be properly addressed to apply a correction for the third dimension in the simulation. First, the integrated electron density in the dot region appears to be a continuous quantity for a two-dimensional system, while there is actually an integer number of electrons n in the dot. If we consider the dot to have a length of 30 nm in the longitudinal direction, we can reasonably assume that the n electrons in the dot correspond to an integrated electron density of $n/30\,\mathrm{nm}^{-1}=0.33n\times10^6\ \mathrm{cm}^{-1}$.

Second, the presence of drain and source contacts affects the electric field under the dot. For instance, when drain and source are kept at 0 V, the electric field is not parallel to the transverse plane, but bends toward the contacts. In other words, drain and source screen the electric field and reduce the sensitivity of the electron density in the channel to the gate voltage. The main consequence of this screening is a poor subthreshold behavior, a problem typical of ultrashort channel metal–oxide–semiconductor field effect transistors (MOSFETs) that also affects this device, as can be appreciated from the results shown in Fig. 2 of Ref. 5.

We take this effect into account through a phenomenological screening capacitance, which is modeled by putting a uniformly distributed positive charge density proportional to the gate voltage V_{gs}, $\rho_{\mathrm{screen}}V_{gs}$, into the channel. The positive charge reduces the electric field in the channel and makes the density of electrons less sensitive to the gate voltage. Since the subthreshold voltage swing ΔS is extremely sensitive to the value of ρ_{screen}, it is important to obtain a reasonable value of ΔS from other considerations, and then to choose a value of ρ_{screen} that provides the best fit for ΔS for our simulation.

In order to determine ΔS, we have performed a semiclassical simulation on the longitudinal section of the device, where the drain and source contacts are visible and therefore their screening effect does not need to be included *ad hoc*. The result of this calculation is shown in Fig. 2, where the integrated electron density in the channel is plotted as a function of the gate voltage for zero electrons in the dot, while keeping source, drain and substrate fixed at 0 V. The subthreshold voltage swing ΔS is then determined by the slope of this function giving $\Delta S\approx0.3$ V/decade.

A corresponding quantum simulation along the longitudinal section would be difficult to achieve, since the channel is not a closed region, and the drain and source contacts contain a prohibitive number of electron states for quantum mechanical simulation.

We choose a screening charge density of $\rho_{\mathrm{screen}}=1.2\times10^{17}\ \mathrm{cm}^3/\mathrm{V}$ for the transverse section, which corresponds to a subthreshold voltage swing very close to 0.3 V/decade, and the total screening capacitance, obtained by integrating $q\rho_{\mathrm{screen}}$ over the channel volume, is 5.49 aF.

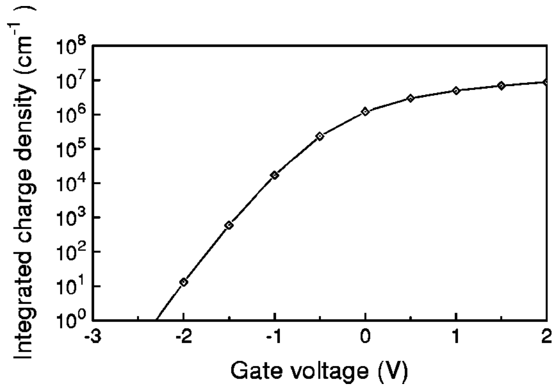

FIG. 2. Integrated electron density in the channel as a function of the gate voltage for zero electrons in the dot, obtained from a semiclassical simulation on the longitudinal section. Source, drain and substrate were held fixed at 0 V.

III. RESULTS

In Fig. 3 the results of the semiclassical simulation are shown. The integrated electron density in the channel is plotted as a function of the gate voltage for zero, one, and ten electrons in the dot.

Since the electrons in the dot have a purely electrostatic effect on the channel, it is sufficient to assume a uniform electron density within the dot: The volume of the dot is $V_{\rm dot}=30\times30\times9\ \ {\rm nm}^3=8.1\times10^{-18}\ \ {\rm cm}^3$, therefore n electrons correspond to a uniform density equal to $n/V_{\rm dot}=1.368\times n\times10^{17}\ {\rm cm}^{-3}$. This approximation seems reasonable, since the second order corrections due to the correct shape of the electron distribution should be smaller than those introduced by the insufficiently known shape and dimensions of the dot. Also, the detailed calculation of the electron density in the dot would not be straightforward, since electrons in the dot are not in equilibrium with the rest of the system, but obey a local equilibrium Gibbs distribution.

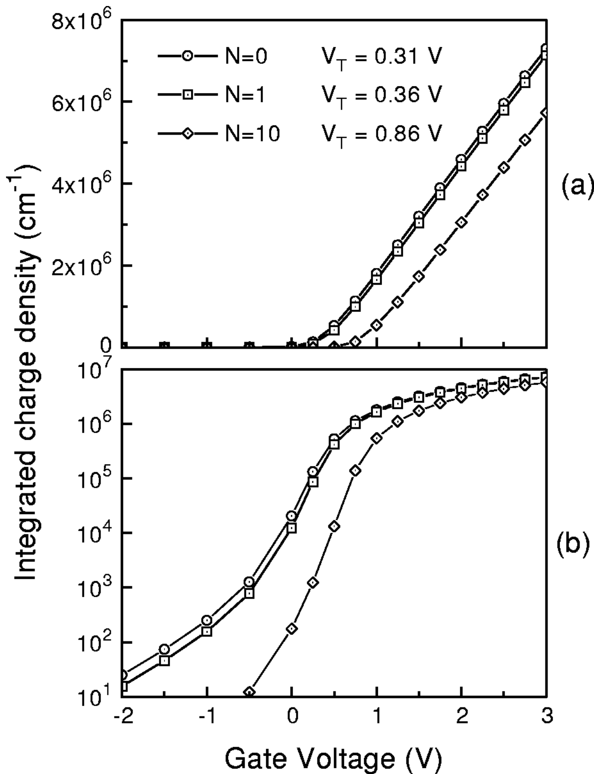

FIG. 3. Integrated electron density in the channel as a function of the gate voltage for zero, one, and ten electrons in the dot, on a linear (a) and a logarithmic scale (b) as obtained from a classical simulation at a temperature of 300 K.

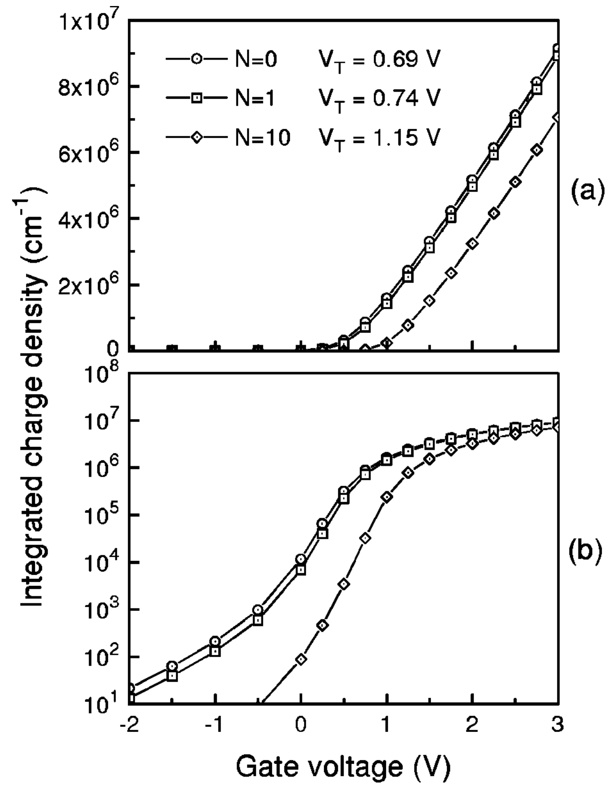

FIG. 4. Integrated electron density in the channel as a function of the gate voltage for zero, one, and ten electrons in the dot, on a linear (a) and a logarithmic scale (b) as obtained from a quantum-mechanical simulation in the channel at a temperature of 300 K.

The threshold voltage V_T and the inversion-layer capacitance $C_{\rm inv}$ between the two-dimensional electron gas (2DEG) and the control gate (i.e., the partial derivative of the total electron charge in the channel to the gate voltage) can be directly obtained from the linear plot [Fig. 3(a)] by extrapolating the linear behavior characteristic of the inversion region (the intercept with the horizontal axis is V_T, while the slope is proportional to the capacitance $C_{\rm inv}$). V_T is 0.31 V when zero electrons are in the dot, and increases by about 55 mV per electron (when one electron is in the dot, V_T is 0.36 V, when ten electrons are in the dot, V_T is 0.86 V). $C_{\rm inv}$ is found to be 1.31 aF.

From the logarithmic plot [Fig. 3(b)], one can see the subthreshold behavior more clearly. At weak inversion, with an applied gate voltage of 0.5 V, ten electrons in the dot imply a 50 times decrease of the integrated charge density, leading to a similar decrease of the current between drain and source.

The corresponding results when the density of electrons in the channel is obtained as a solution of the Schrödinger equation are shown in Fig. 4 and are similar to the semiclas-

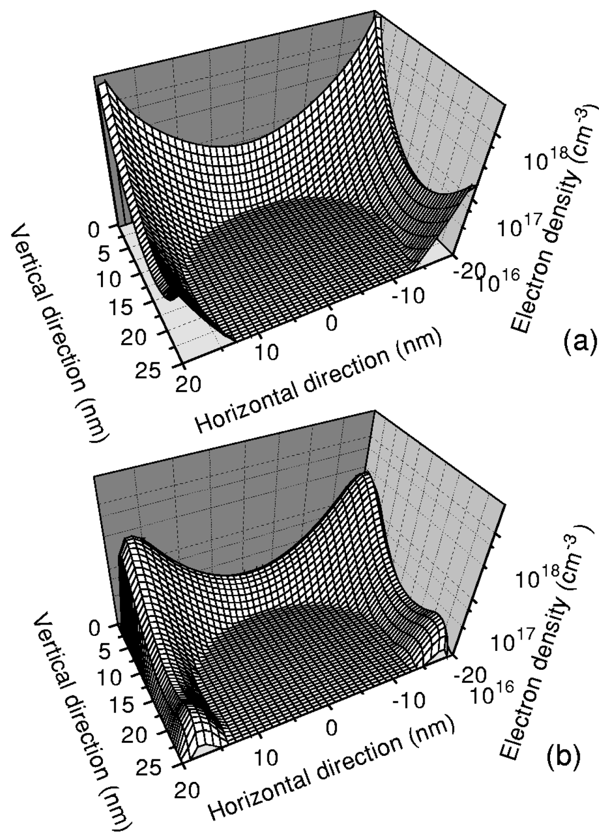

FIG. 5. Surface plots of the electron density in the channel for a gate voltage of 0.5 V and no electrons in the dot for a classical (a) and a quantum (b) simulation.

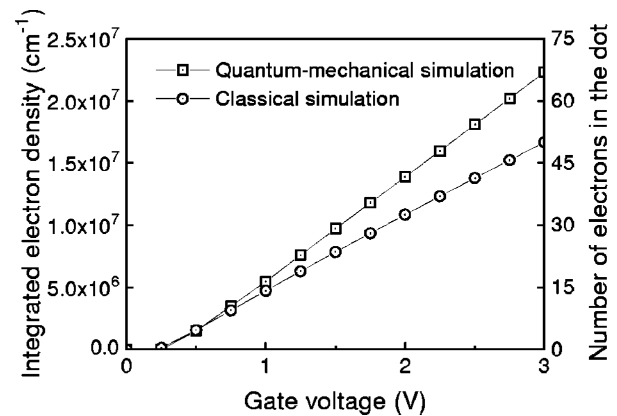

FIG. 6. Integrated electron density in the dot region (left axis) and corresponding number of electrons in the dot (right axis) as a function of the gate voltage at a temperature of 4.2 K, as obtained from a classical and a quantum-mechanical simulation.

sical results. However, quantum confinement effects in the channel are quantitatively relevant: when zero electrons are in the dot V_T is found to be 0.69 V, a substantial increase with respect to the value of 0.31 V obtained from the semiclassical calculation. The extra voltage is needed to overcome the confinement energy of electrons in the inversion layer, while second order terms, such as exchange and correlation, are only negligible corrections to the confining potential. A few estimates of the effects of quantum confinement in the inversion layer on the threshold voltage can be found in the literature:[7–9] while there is a qualitative agreement with our results, it seems that in our case the quantum correction to the threshold voltage is larger by a factor from 2 to 10. Actually, the difference could be due to the fact that the inversion layer of our structure is finite and nonplanar, while all results found in the literature refer to a 2DEG on an infinite plane. This consideration is also based on the fact that the simulation performed on an infinitely long structure with the same transversal section as ours gives a quantum correction to the threshold voltage of about 30 mV.

The increase of V_T per electron is 46 mV, reduced with respect to the semiclassical case, while the inversion-layer capacitance in this case is 1.85 aF, higher than the semiclassical value. Again, these corrections are mainly due to the nonplanarity of the 2DEG in the channel, since for an infinite two-dimensional electron gas the inversion-layer capacitance would be decreased, due to the so-called quantum capacitance effect.[10] Figures 5(a) and 5(b) show the electron densities resulting from the classical and the quantum-mechanical simulation. We see that in both cases electrons are mainly accumulated at the top edges of the channel, where the field is higher, while the lateral sides of the channel are rather far from inversion.

Also, as expected, the shape of the electron density is much smoother in the quantum simulation than in the classical case. The result is a more uniform charge distribution at the $Si-SiO_2$ interface, yielding a higher effective area of the inversion capacitor. This effect overcompensates the quantum capacitance effect, and the inversion capacitance in the quantum case is higher than in the semiclassical one.

For comparison, a simplistic calculation of the capacitance between the control gate and the channel, done by assuming that the total capacitance is the sum of that of two lateral parallel plate capacitors of area $25\times 30\ nm^2$ with a 18 nm SiO_2 dielectric and a top parallel plate capacitor of area $30\times 30\ nm^2$, and a dielectric made of three different layers (9 nm SiO_2, 8 nm Si, 2.5 nm SiO_2), would give $C_{inv} = 5.05$ aF. The difference with the values previously calculated is mainly due to the screening charge density, which takes into account the capacitive coupling with the source and drain electrodes making the charge in the inversion layer less sensitive to the control gate voltage.

In order to investigate the charging process in the floating dot, we have performed both a quantum and a classical simulation of the complete dot-channel system at a temperature of 4.2 K. In this case, the charge in the dot is not imposed as fixed uniform density, but is self-consistently calculated under the assumption that after the gate voltage is applied, enough time is passed to let the dot be in equilibrium with the rest of the device. The evaluation and the optimization of charging and decharging times is a crucial issue towards the applicability of quantum-dot flash memories, but is beyond the scope of this paper and will be addressed separately.

In Fig. 6 the integrated electron density in the dot is plotted as a function of the gate voltage, as calculated from the classical and quantum-mechanical model at 4.2 K. The number of electrons corresponding to the integrated electron

density is plotted on the right axis. We find that a 57 mV voltage shift is required to put a classical electron in the dot, while a voltage shift of 40 mV is sufficient in the quantum-mechanical case. The smaller voltage shift for the quantum case is not necessarily intuitive and shows that a careful evaluation of the quantum effects in capacitive couplings is needed. However, both the classical and the quantum result are consistent with a simple calculation based on a parallel plate approximation of the capacitance C_{gd} between gate and dot, yielding $C_{gd}=3.45$ aF. The shift in the gate voltage required to put an extra electron in the dot would be in this case $q/C_{gd}=46$ mV.

IV. DISCUSSION

We have shown that the essential electrical features of quantum-dot flash memories can be obtained from a self-consistent solution of the two-dimensional Poisson and Schrödinger equation in the transverse section, provided that three-dimensional effects are properly taken into account. For quantitative simulation of experimental devices, the knowledge of such parameters as stray and interface charges and the details of the actual geometries seem to be more important than a complete three-dimensional simulation.

These simulations may be helpful in the interpretation of experimental results. For instance, the device described by Welser *et al.*,[5] which we took as a model for our calculation, is indicated to exhibit a 0.75 V shift in the threshold voltage V_T when about seven electrons are in the dot. These seven electrons would be put in place by slowly sweeping the gate voltage over about 6 V. This interpretation is in conflict with the results of our simulation, and also with the simple calculation mentioned above. Our results indicate that only about 40 mV are needed to add an extra electron to the dot; more than 60 electrons would be added with a 6 V sweep. In this regard, we think that the slight bumps in the plot of Fig. 3 in Ref. 5 are not a direct measure of the actual number of electrons in the dot.

Assuming a population of about 60 electrons in the dot, the measured shift in the threshold voltage V_T of 0.75 V would result from a threshold voltage shift of approximately 12.5 mV per electron in the dot. This value is of course much smaller than the 46 mV voltage shift per electron we have calculated, but one must consider that the actual capacitance between the gate and the channel is difficult to determine exactly and is very likely higher in the experimental structure than assumed in our calculation. This is due to the shorter distance between the gate and the lower part of the channel [as can be seen in Fig. 1(c) of Ref. 5], caused by the difficulty in controlling the boundaries during fabrication. Therefore, in the experimental device, the lower edges of the channel could be much closer to inversion than in our simulated device, making V_T much less dependent on the charge in the dot or, in other words, the effective area of the capacitor between the gate and the channel would be much larger in the experimental device. However, our result concerning the number of electrons in the dot should not be much affected by this geometric effect, since it only depends on the capacitance between the control gate and the dot.

As a final remark, we would like to stress that, in addition to the classical capacitance between gate and channel considered in Refs. 3 and 4, the screening effects of the source and drain contacts together with the charge distribution in the channel also have an important influence on the calculated threshold-voltage shift per electron in the dot. The reason for this is that only the interface region in strong inversion would screen effectively the field as an ideal conductor, but, as Fig. 5(b) shows, strong inversion is realized only in a portion of the dielectric-channel interface. Furthermore, the density of states calculated from the self-consistent solution of Schrödinger's and Poisson's equation provides quantum corrections for the inversion-layer capacitance and threshold voltage of finite and nonplanar structures that are very different from those obtained by extrapolating results for infinite planar structures.

ACKNOWLEDGMENTS

This work has been supported by the Ministry for the University and Scientific and Technological Research of Italy, by the Italian National Research Council (CNR), by a NATO Collaborative Research Grant No. CRG.950753, and by the National Science Foundation through Grants Nos. ECS 95-26127 and ECS 95-09751. The authors also acknowledge the support (A.T.) of the CSE fellowship program of the University of Illinois.

[1] S. Tiwari, F. Rana, H. Hanafi, A. Hartstein, and E. F. Crabbé, Appl. Phys. Lett. **68**, 1377 (1996).
[2] S. Tiwari, F. Rana, K. Chan, L. Shi, and H. Hanafi, Appl. Phys. Lett. **69**, 1232 (1996).
[3] L. Guo, E. Leobandung, and S. Y. Chou, Appl. Phys. Lett. **70**, 851 (1997).
[4] A. Nakajima, T. Futatsugi, K. Kosemura, T. Fukano, and N. Yokoyama, Appl. Phys. Lett. **70**, 1742 (1997).
[5] J. J. Welser, S. Tiwari, S Rishton, K. Y. Lee, and Y. Lee, IEEE Electron Device Lett. **18**, 278 (1997).
[6] A. Trellakis, A. T. Galick, A. Pacelli, and U. Ravaioli, J. Appl. Phys. **81**, 7880 (1997).
[7] S. Takagi and A. Toriumi, IEEE Trans. Electron Devices **42**, 2125 (1995).
[8] F. Rana, S. Tiwari, and D. A. Buchanan, Appl. Phys. Lett. **69**, 1104 (1996).
[9] G. Chindalore, S. A. Hareland, S. Jallepalli, A. F. Tasch, Jr., C. M. Maziar, V. K. F. Chia, and S. Smith, IEEE Electron Device Lett. **18**, 206 (1997).
[10] S. Luryi, Appl. Phys. Lett. **52**, 501 (1988).