Self-Consistent Simulation of Quantization Effects and Tunneling Current in Ultra-Thin Gate Oxide MOS Devices

A. Ghetti[a], A. Hamad[b], P.J. Silverman[a], H. Vaidya[b] and N. Zhao[c]
a) Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974, USA
b) Bell Laboratories, Lucent Technologies, Orlando, FL 32819, USA
c) CIRENT Semiconductor, Orlando, FL

***Abstract*—In this paper we report on the self-consistent modeling and simulation of quantization effects and tunneling current in MOS devices. The simulation model features an original scheme for the self-consistent solution of Poisson and Schrödinger equations and it is used for the extraction of the oxide thickness, by fitting CV curves, and the calculation of the tunneling current. Simulations and experiments are compared for different device types and oxide thicknesses (1.5-6.5nm) showing good agreement and pointing out the importance of quantum mechanical modeling and the presence of many tunneling mechanisms in ultra-thin oxide MOS devices.**

I. Introduction

Sub-0.1μm technologies require gate oxide thicknesses $t_{OX} < 3nm$ [1]. For such thin oxides, a significant tunneling current is expected even at normal operating conditions and quantum mechanical (QM) effects plays an important role. In addition, matching simulations and experiments can be exploited for electrical characterization and physical parameter extraction (especially t_{OX}). In order to accomplish these goals an accurate and self-consistent modeling of tunneling current and quantization effects is needed, and the absence of device dependent parameters is a desirable condition.

II. Simulation Model

The present model refers to a 1D polysilicon-oxide-silicon structure. A many valley ellipsoidal parabolic band for silicon electrons (m_{te}=0.98m_0, m_{te}=0.19m_0) and holes (m_{lh}=0.16m_0, m_{hh}=0.5m_0), and a spheric parabolic band for oxide are assumed. The poly is modeled as silicon but considering the appropriate work function accounting for the correct doping level and band gap narrowing [2]. Thermal equilibrium is assumed, then Fermi-Dirac statistics is adopted. The self-consistent solution of Poisson and Schrödinger equations is achieved through an original iteration scheme. The Poisson equation is written in the following non linear form:

$$-\nabla \cdot (\epsilon \nabla V^{k+1}) =$$

$$q\Big[N_V F_{\frac{1}{2}}\left(\frac{q}{k_B T}(V^k - V^{k+1}) + F^{-1}_{\frac{1}{2}}\left(\frac{p}{N_V}\right)\right) - N_C F_{\frac{1}{2}}\left(\frac{q}{k_B T}(V^{k+1} - V^k) + F^{-1}_{\frac{1}{2}}\left(\frac{n}{N_C}\right)\right) - N(z)\Big], \quad (1)$$

where $F_{\frac{1}{2}}(\eta)$ is the Fermi-Dirac integral of order $\frac{1}{2}$ and $N(z)$ is the net doping. The particular form of (1) reduces to the usual linear expression of the Poisson equation near convergence ($V^{k+1} = V^k$), but helps to reduce the numerical instability typical of self-consistent Poisson-Schrödinger iterative solution, thus speeding up convergence. For a single bias point the number of iteration needed to converge is typically 5. The heaviest CPU burden is the computation of the Fermi-Dirac integral that can be reduced adopting a spline interpolation.

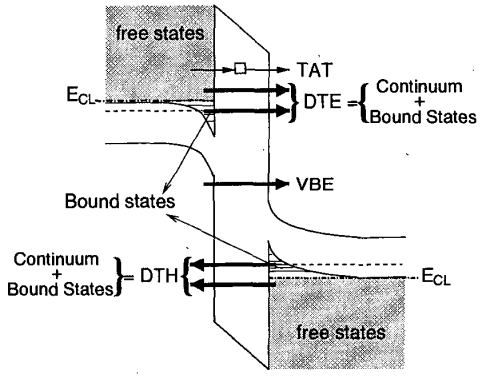

Fig. 1. Schematic representation of the different components of the tunneling current considered in this work, and of the different types of $\vec{k}$ states. TAT is computed accounting for the contribution of both free and bound states. Electrons/holes with energy above/below E_{cl} form a free gas.

In order to compute the correct charge density in both accumulation and inversion layers we consider bound states up to a given threshold energy E_{cl}, above which carriers are thought to form a free gas [3] (Fig. 1). Thus:

$$n(z) = n_{3D}(z) + n_{2D}(z)$$

$$n_{2D}(z) = \sum_{ij}^{E_{ij}<E_{cl}} \frac{g_j m_{d_j} k_B T}{\pi\hbar^2} \ln\left(\frac{1+e^{\frac{E_F-E_{ij}}{k_BT}}}{1+e^{\frac{E_F-E_{cl}}{k_BT}}}\right) |\zeta_{ij}(z)|^2$$

$$E_M(z) = max(E_{cl}, E_C(z))$$

$$n_{3D}(z) = N_C F^{(i)}_{\frac{1}{2}}\left(\frac{E_F - E_M(z)}{k_BT}, \frac{E_{cl} - E_C(z)}{k_BT}\right),$$

where $E_C(z)$ is the conduction band edge, g_j is the degeneracy of the $j-th$ valley, m_{d_j} its density of states effective mass, E_{ij} and ζ_{ij} are respectively the energy level and the corresponding envelope wave function of the $i-th$ bound state in the $j-th$ valley, and $F^{(i)}_{\frac{1}{2}}(z,b)$ is the incomplete Fermi-Dirac integral as defined in [4]. Hole quantization is treated in a symmetric way.

This approach gives the correct continuum and 2D density of states accounting only for the lowest quantum levels and simplifies the choice of boundary conditions for the Schrödinger equation. Quantum levels are computed using Sturm sequencing and bisection, and envelope functions are found by inverse iteration. Gate depletion effects are implicitly accounted for by solving for the potential also over the gate region. Capacitance is computed differentiating charge-to-voltage characteristics.

The self-consistent potential profile is then used to compute the transmission probability (T) through an exact solution of the Schrödinger equation in terms of Airy's functions [5]. Barrier lowering is not included, while parallel momentum conservation is enforced. The tunneling current is found integrating the contribution of all $\vec{k}$ states. Because of the assumed silicon band structure, the sum over all free states reduces to:

$$J_{3D} = \frac{qgm_d}{4\pi^3\hbar^3} \int_{E_{cl}}^{\infty} dE_\perp \int_0^{\infty} dE_{||}\ \Delta f(E_\perp + E_{||}) \int_0^{2\pi} T(E_\perp, E_{||}, \theta) d\theta, \qquad (2)$$

where Δf is the difference of the Fermi-Dirac statistics at the two ends of the barrier. The lifetime of a bound state is given by [6]:

$$\tau_L = \frac{\tau(E_i)}{T} = \frac{\int_a^b \sqrt{\frac{2m}{E_i - E_C(z)}}\, dz}{T},$$

where a and b are the classical turning points. Thus the contribution of the bound states is given by

$$J_{2D} = \sum_i^{E_i<E_{cl}} \frac{qg_i m_{d_i}}{2\pi^2\hbar^2\tau(E_i)} \int_0^{\infty} dE_{||}\ \Delta f(E_i + E_{||}) \int_0^{2\pi} T(E_i, E_{||}, \theta) d\theta \qquad (3)$$

There is no conceptual difference between conduction band and valence band electrons, apart from their different band structure. Therefore the valence to conduction band tunneling current can be computed as in (2) taking in to account the valence band structure and the appropriate potential profile.

Trap-assisted electron tunneling is also included in the simulation. It is modeled according to an elastic two step model similar to [7]. In steady state condition, the current flowing via the traps located at z is given by the balance between the tunnel-in current ($J_{in}(z)$) and the tunnel-out current ($J_{out}(z)$). Then

$$J_{TAT} = \int_0^{t_{OX}} \sigma\, N_{tr}(z)\, \frac{J_{in}(z)\, J_{out}(z)}{J_{in}(z) + J_{out}(z)}\, dz, \qquad (4)$$

where σ is the trap capture cross section and N_{tr} is the trap concentration. J_{in} and J_{out} are computed according to (2) or (3) depending on the initial state.

In summary the model computes the following tunneling current components (Fig. 1): conduction band electrons (DTE) and valence band holes (DTH) tunneling from bound and free states, valence band electrons (VBE) tunneling, and trap-assisted electron tunneling (TAT).

This model has been implemented in the simulation program called QUASI.

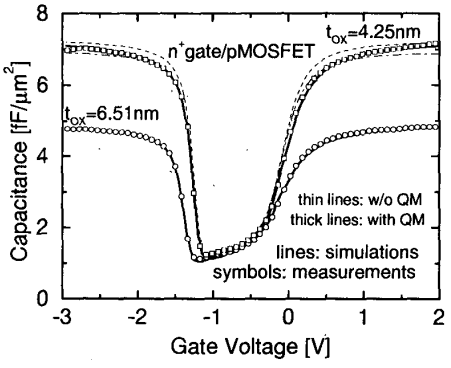

Fig. 2. Quasi static CV measurements (symbols) and simulations performed with (thick solid lines) and without (thin lines) the inclusion QM effects for n⁺poly pMOS transistors. t_{OX} given by the best QM fitting is indicated along each curve. Classical simulations fitting different parts of the experimental data provide $t_{OX} = 4.4nm$ (dashed thin line) and $t_{OX} = 4.6nm$ (dot-dashed thin line).

III. Model Calibration

For a known structure, i.e. potential profile, DTE computation depends only on the oxide effective mass (m_{OX}) and barrier height (Φ_B). Since there is a substantial agreement in the literature on their numerical value

($m_{OX} = 0.5m_0$ and $\Phi_B = 3.1eV$), the DTE calculation does not rely on device dependent parameters. Therefore we first verified that this model is able to reproduce the tunneling current with an *independent* determination of the potential profile.

To this purpose, we first analyzed a set of relatively thick oxide devices because of the availability of both CV and IV measurements. First, physical t_{OX} was extracted by fitting experimental CV data with QM simulations. Fig. 2 reports CV measurements and simulations of n$^+$poly pMOS transistors. The excellent agreement in both accumulation and inversion regimes is due to the inclusion of QM effects for both electrons and holes and the adoption of the Fermi-Dirac statistics. Without QM effects (classical case), simulations don't match data as well and provide a larger t_{OX}. Then, tunneling IV characteristics were simulated *with the same parameters* (Fig. 3).

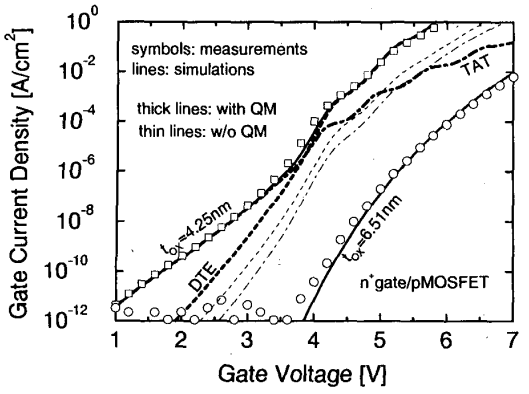

Fig. 3. Tunneling IV measurements (symbols) and simulations (lines) of the same devices of Fig. 2 in accumulation regime. Quantum mechanical simulations (thick lines) featuring t_{OX} given by the best QM fit of Fig. 2 are in good agreement with experiments. Classical DTE simulations featuring $t_{OX} = 4.4nm$ (dashed line) and $t_{OX} = 4.6nm$ (dot-dashed line) underestimate the tunneling current. The thinner device shows a native trap assisted component (TAT). The best fit was achieved with $\sigma N_{tr} = 4.5cm^{-1}$.

In the case of the thickest oxide, DTE fits very well the experiments. For the 4.25nm device, DTE simulated including QM effects fits experimental data at high enough voltages, also reproducing the characteristic oscillations due to QM reflection at the oxide/anode interface, while classical simulations provide a smaller tunneling current of more that 1 order of magnitude.

The extra current with respect to DTE featured by experiments at low voltages is a TAT component due to *native traps* [8]. It was computed assuming a trap distribution uniform in space and energy, and using the product of the trap cross section σ and trap density N_{tr} as a fitting parameter.

IV. Simulation of Ultra-Thin Oxides

Figs. 4, 5 compare experimental data with simulations of n$^+$poly nMOS transistors featuring t_{OX} in the range 1.5-6.5nm, in inversion and accumulation regime respectively. Here t_{OX} was used as a fitting parameter because of the impossibility to take reliable CV measurements of such small area devices. The adopted values of t_{OX} are compared to the average value of ellipsometric measurements in the caption of Fig. 4. They are well inside the spread of the t_{OX} distribution due to the non uniformity of the thin oxide across the wafer.

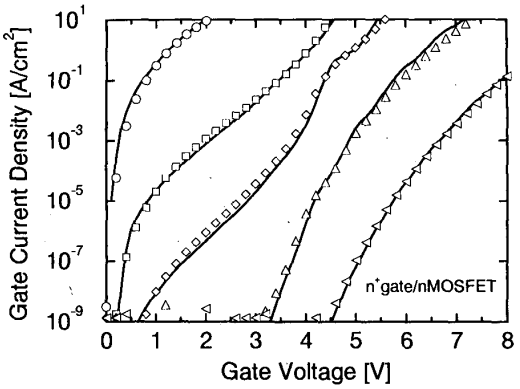

Fig. 4. Tunneling IV measurements (symbols) and simulations (lines) for n$^+$poly nMOS transistors in inversion regime. Matching simulations with experiments resulted in the following values for t_{OX} (in nm) for the different curves (in parenthesis the corresponding value from ellipsometric measurements); from left to right: 1.55(1.56), 2.55(2.47), 3.35(3.27), 4.64(4.59), 6.64(6.56). Devices with t_{OX} of 2.47nm and 3.27nm feature a TAT component at low gate voltage as explained in [8].

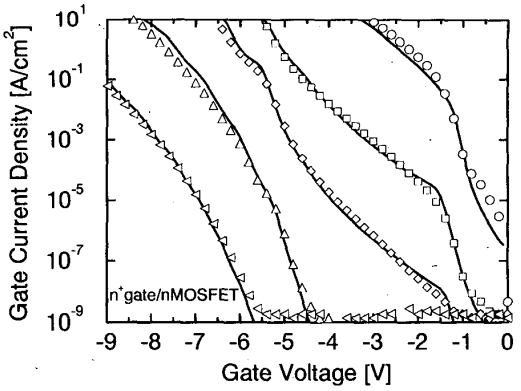

Fig. 5. Tunneling IV measurements (symbols) and simulations (lines) for the same devices, symbols and simulation parameters of Fig. 4, but in accumulation regime.

If the oxide is thin enough, it is possible to mea-

sure the valence band electron tunneling current component (VBE) by biasing in inversion a nMOS transistor and measuring the substrate current (I_B), that, in this case, is made of the holes left behind by the valence band tunneling electrons [9]. Fig. 6 compares I_G and I_B measurements to simulations for the same device with $t_{OX} = 2.47nm$ in Figs. 4,5. VBE simulation agrees pretty well with I_B. It was computed using $m_{OX} = 0.42m_0$.

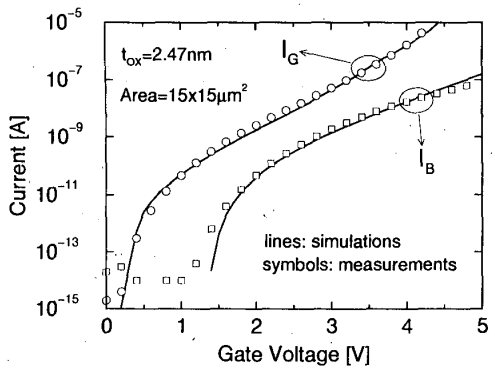

Fig. 6. I_G and I_B measurements (symbols) and simulations (lines) for the device with $t_{OX} = 2.47nm$ of Fig. 4. The simulation of I_G accounts for both DTE and TAT components, while VBE simulation coincides with I_B.

As last example, Fig. 7 reports experiments and simulations of p^+gate over lightly doped ($10^{15}cm^{-3}$) p-substrate capacitors with t_{OX} in the range 1.5-2nm. t_{OX} needed to fit experiments are within 7% of the ellipsometric measurements. In this case, gate current is the sum of DTH and VBE components. Fig. 8 shows a more detailed analysis of the gate current components for the 2nm device of Fig. 7. For $-1.5V < V_G < 0V$, DTH (computed assuming $m_{OX} = 0.42m_0$ and $\Phi_B = 4.8eV$) dominates the conduction, while VBE will eventually take over when the gate valence band faces the substrate conduction band. Notice that the exclusion of QM effects leads to an underestimation of DTH, pointing out the importance of accounting for QM effects in the calculation of tunneling current even for such low doping.

V. Conclusion

In summary, we have presented a self-consistent simulation model of quantization effects and tunneling current that does not rely on device dependent parameters. It provides a very good agreement with experimental CV and tunneling IV data over a broad variety of device types and t_{OX}, provided that an accurate modeling of quantization effects and of all tunneling mechanisms is included in the simulation.

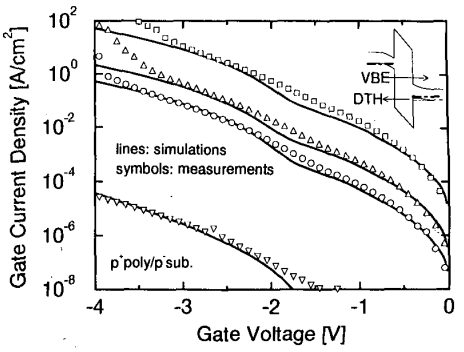

Fig. 7. Tunneling IV measurements (symbols) and simulations (lines) for p^+poly over p-substrate MOS capacitors. Ellipsometric measurements of t_{OX} (from top to bottom): 1.53nm, 1.86nm, 2nm, 2.95nm.

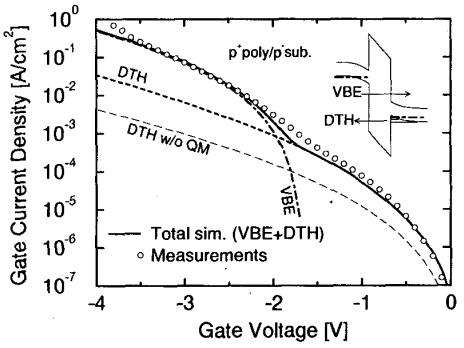

Fig. 8. Analysis of the tunneling current components for a p^+-poly over p-type substrate MOS capacitor with an estimated $t_{OX} = 2nm$. Measured (o) and simulated (solid line) gate tunneling current are reported together with the simulated DTH (dashed line) and VBE (dot-dashed) components. The thin long dashed line represents the simulated DTH component if hole quantization is neglected.

Acknowledgements

The authors would like to thank C.T. Liu and J. Bude for many helpful discussions, and G. Timp and T.W. Sorsch for providing some of the devices tested.

References

[1] C.T. Liu, in *IEDM Technical Digest*, p. 747, 1998.

[2] N. Lifshitz, *IEEE Trans. on Electron Devices*, vol. 32, no. 3, p. 617, 1985.

[3] C. Bowen *et al.*, in *IEDM Technical Digest*, p. 869, 1997.

[4] M. Goano, *Solid State Electronics*, vol. 36, no. 2, p. 217, 1993.

[5] W. Lui *et al.*, *Journal of Applied Physics*, vol. 60, no. 5, p. 1555, 1986.

[6] F. Rana *et al.*, *Applied Physics Letters*, vol. 69, no. 8, p. 1104, 1996.

[7] E. Suzuki *et al.*, *Journal of Applied Physics*, vol. 60, no. 10, p. 3616, 1986.

[8] A. Ghetti *et al.*, *to be published in Proc. INFOS*, 1999.

[9] B. Eitan *et al.*, *Applied Physics Letters*, vol. 43, p. 106, 1983.