Physical Modeling of Deep-Submicron Devices

Andreas Schenk
Integrated Systems Laboratory,
ETH Zürich,
Gloriastr. 35, CH-8092 Zürich
Tel. +41 1 632 66 89
schenk@iis.ee.ethz.ch

Abstract

This paper gives an overview of established methods to describe quantum effects in deep-submicron CMOS. Recent progress in the integration of quantum models in TCAD packages is illustrated by a number of applications.

1. Introduction

Today's challenges in the field of modeling and simulation of deep-submicron devices ($\mathrm{L_{eff}} \leq 100\,\mathrm{nm}, \mathrm{t_{ox}} \leq 3\,\mathrm{nm}$) are closely linked to the goals and uncertainties of the SIA Roadmap. They include (i) effects of quasi-ballistic transport (prediction of on-current, role of hot-carriers in gate currents and interface-trap generation), (ii) quantum effects (confinement, gate tunneling leakage, GIDL, channel mobility, source-to-drain tunneling), (iii) generation of interface-traps, (iv) spatial and temporal fluctuations (discreteness of doping, RF and 1/f noise), (v) long-range electron-electron interaction, and (vi) transport modeling for new materials (strained silicon, SiGe). This paper will focus on the simulation of quantum effects in state-of-the-art TCAD. Various quantization models are compared with each other in terms of accuracy and feasibility. Whereas in CMOS quantum effects are undesirable and TCAD is being used to minimize their impact on device performance, the post-CMOS era could utilize quantum effects for certain applications.

2. Quantization models

For the inclusion of quantization effects in a classical device simulator, a straightforward approach is to introduce a "quantum potential" Λ in the classical formula of the electron density (expressions for holes suppressed throughout):

$$n = N_c \exp\left[\beta\left(E_{F,n} - E_c - \Lambda\right)\right] \ , \quad (1)$$

($\beta = 1/kT$). When Fermi statistics is required, the exponential function is replaced by the Fermi integral of order 1/2. Obviously, only quantum effects related to density modifications can be described by Eq. (1).

2.1. van Dort model

The van Dort quantum correction model [1] computes Λ as a function of the (local) electric field $F_\perp$ normal to the semiconductor-insulator interface:

$$\Lambda = a_{\mathrm{fit}}\, h(\mathbf{d})\, (\beta\epsilon_0\epsilon_s/4)^{1/3}\, |F_\perp|^{2/3} \ , \quad (2)$$

where $a_{\rm fit}$ is a fit factor and $h(\mathbf{d})$ a turn-off function which restricts the model to a near-interface region. Inserted in Eq. (1), Λ acts as an effective band gap widening. Depending on the sign of $F_\perp$, it is either applied to electrons (E_c shift) or to holes (E_v shift). Since the model is based on the expression for the lowest eigenenergy of a carrier in a triangular potential well, it is suited for simulations of bulk MOSFETs and, to some extent, also of SOI and double-gate transistors (provided the body thickness does not reach quantum-mechanical length scales). Although the channel density distribution in bulk MOSFETs is not reproduced (see Fig. 1), important terminal characteristics are often well-described. The model is numerically robust and fast.

2.2. 1D Schrödinger-Poisson solver

The 1D Schrödinger equation is the physically most sound quantization model. Here, the quantum potential Λ follows by equating the density (1) with the expression

$$n(z) = \frac{1}{\beta\pi\hbar^2}\sum_{j,\nu}\left|\Psi_j^{(\nu)}(z)\right|^2 m_{xy}^{(\nu)}(z)\times \\ \times\exp\left[\beta\left(E_{F,n}(z)-E_j^{(\nu)}\right)\right]\ , \quad (3)$$

where $\Psi_j^{(\nu)}$ and $E_j^{(\nu)}$ are the j-th eigenfunction and eigenenergy for valley ν obtained by numerical solution of the 1D (effective mass) Schrödinger equation in z-direction (typically, the direction perpendicular to the Si-SiO_2 interface) [2, 11, 4]. The boundary conditions at the ends of the domain $[z_-, z_+]$ (defining the total 'quantum box')

$$\Psi_j^{(\nu)\prime}/\Psi_j^{(\nu)} = \pm\sqrt{2m_z^{(\nu)}|E_j^{(\nu)}-E_c|}/\hbar\ , \quad (4)$$

based on a WKB argument, were found to be superior over Dirichlet boundary conditions [3, 4]. However, especially under flat-band conditions, where many continuum states are involved, results are sensitive to the extension of the quantum box in non-barrier regions.

The CPU time depends on the number of grid lines in quantization direction (and hence may be considerable), and convergence problems occur at large drain currents. A tensor-product grid is needed in the quantum boxes (typically, gate oxide and channel). These drawbacks make the method more suited for calibration and validation purposes than for optimization.

2.3. Density gradient model

In the density gradient (DG) model (or 'quantum drift-diffusion' (QDD)) [5, 6, 7, 3, 4] Λ is given by the PDE:

$$\Lambda = -\gamma\frac{\hbar^2}{12\,m}\left[\nabla^2\log n+\frac{1}{2}\left(\nabla\log n\right)^2\right] \\ = -\gamma\frac{\hbar^2}{6m}\frac{\nabla^2\sqrt{n}}{\sqrt{n}}\ , \quad (5)$$

where γ is a fit factor. A number of approximations are necessary to obtain Eq. (5), e.g. thermodynamic equilibrium and isotropy of the effective mass m. The implementation in [4] uses the DOS mass for m and generalizes (5) for semiconductor regions to

$$\Lambda = -\gamma\frac{\hbar^2}{12\,m}\left\{\nabla^2\left(\beta E_{F,n}-\beta\bar{\Phi}\right)+ \\ +\frac{1}{2}\left[\nabla\left(\beta E_{F,n}-\beta\bar{\Phi}\right)\right]^2\right\}\ , \quad (6)$$

Here, $\bar{\Phi}$ is the smoothed potential $\bar{\Phi} = E_c + \Phi_m + \Lambda$, containing the electrostatic part E_c (which includes band edge discontinuities), a mass driving term Φ_m (resulting from DOS discontinuities), and the quantum potential Λ, the addition of which does not contribute in lowest-order quantum correction, as long as the Born approximation is justified [8]. In practically relevant cases the

latter is strongly violated, and the form (6) (or Eq. (5) with the *quantum-mechanical* density n) represents a nonperturbative formulation, which deserves further research. The main effect of the quantum potential Λ is to smooth out rapid changes of the potential on a length scale $\sqrt{\gamma\beta\hbar^2/2m}$ ($\approx$ thermal de Broglie wave length).

Discretization of Eq. (6) on unstructured grids is mandatory for professional TCAD and has been demonstrated in [8]. Instead of inserting Λ into the current equation, it is treated as new variable, which increases the number of unknowns of the nonlinear system, but conserves the sparsity structure of the Jacobian. The extended system is solved by a coupled Newton. Thus, the DG model is numerically robust, but convergence is not necessarily faster than for the Schrödinger equation. Another obvious advantage of the DG model is that it is *per se* multidimensional, whereas the 1D-Schrödinger method relies on the adiabatic decoupling of the 3D Schrödinger equation. Extensive simulations and comparisons with the more accurate Schrödinger method have shown that there is a "universal" fit parameter $\gamma = 3.6$ for silicon, a value close to theoretical

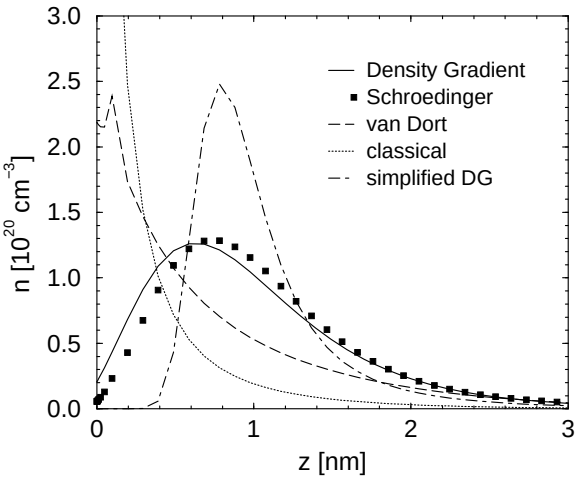

Figure 1. Electron density in different quantization models.

expectations. "Universal" means that γ does not depend on oxide thickness, channel doping, or device temperature. Despite some uncertainties concerning physical rigor, the DG model is promising and worthwhile to be investigated in more detail. Open questions include the boundary conditions, the re-calibration of physical models, the application to thermionic tunneling (e.g. Schottky contacts) and band-to-band tunneling, the mass anisotropy, and others.

3. Reduction of gate capacity and V_T shift

The most familiar quantization effects in CMOS are the reduction of gate capacity and the shift of the threshold voltage. In Fig. 1 the electron density in a MOS-diode with $5 \times 10^{17}\,\mathrm{cm}^{-3}$ channel doping, 4 nm oxide thickness and 4 V gate voltage is shown for the different quantization models discussed in the text. The agreement between 'DG' and 'Schrödinger' can be made per-

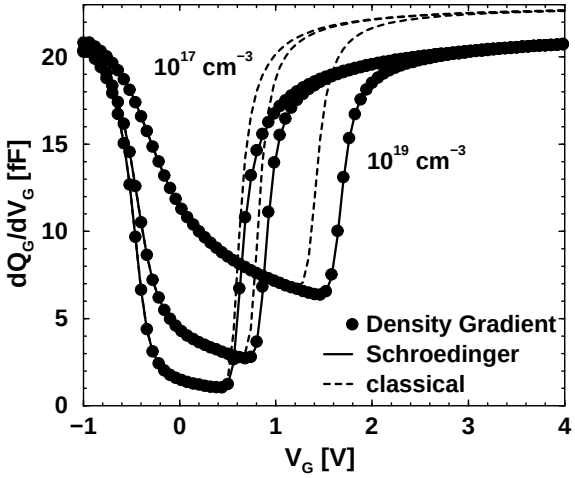

Figure 2. CV curves for different substrate doping.

fect, if $\gamma = 1$ is used in the oxide (instead of 3.6 as for silicon). In the 'simplified density gradient' model the second term in Eq. (6) is neglected ($\gamma = 0.3$ in this

case). Fig. 2 presents CV curves for a MOS diode with 1.5 nm oxide thickness, $1\,\mu\mathrm{m}^2$ area and different doping concentrations in the substrate. The same good agreement between 'Schrödinger' and 'DG' is found for an SOI MOSFET with 5 nm body thickness, 80 nm channel length, 1.5 nm oxide thickness, and 50 mV drain voltage as shown in Fig. 3. As the DG model is sensitive to rapid

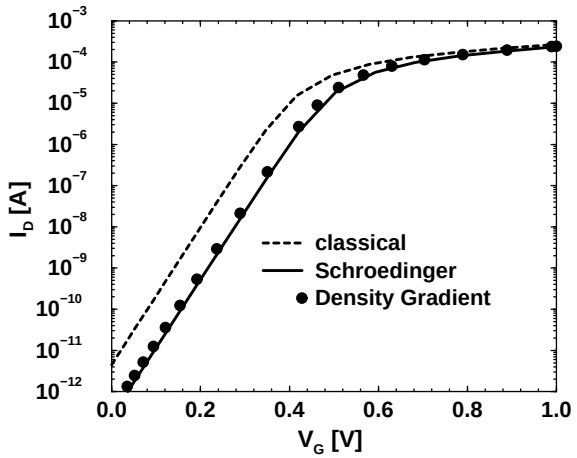

Figure 3. Transfer characteristics of an SOI MOSFET.

changes of the potential *everywhere* in the device, it also gives a quantum correction in the proximity of the poly-insulator interface, if the poly gate is included in the simulation. The carriers are repelled from the potential wall, and the electron density drops towards the interface. Assuming a homogeneous background doping, a thin space-charge layer will form and the (positive) gate charge increases. For lowly doped poly ($< 10^{19}\,\mathrm{cm}^{-3}$) the effect is negligible, however, for realistic doping concentrations it might even over-compensate the threshold voltage shift caused by quantization in the channel (see Fig. 4). This effect has been reported in Ref. [9] based on a solution of the Schrödinger equation and has also been observed using the DG model in Ref. [10]. Fig. 4 presents the low-frequency CV plot

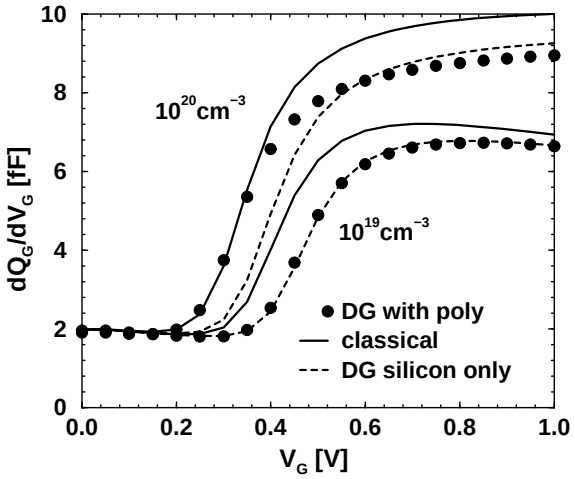

Figure 4. Effect of quantization in poly.

for a NMOS diode with 3 nm oxide thickness, $5 \times 10^{17}\mathrm{cm}^{-3}$ doping concentration, and poly doping levels of $10^{19}\mathrm{cm}^{-3}$ and $10^{20}\mathrm{cm}^{-3}$, obtained by a classical simulation, a calculation with the DG model everywhere in the device, and with the DG model correction in silicon only (but not in poly) [4]. Identical results for poly quantization can be obtained using the self-consistent Schrödinger model [4] (not shown), which proves that this effect is not an artifact of the DG model. In the simulation, poly was treated as silicon, i.e. with the same m and γ as for the channel. Is the effect of quantum-mechanical depletion in the poly-gate real or not? If a 1 nm wide layer is left undoped, the effect almost vanishes [4]. The actual doping level in this ultra-thin layer is not accessible by SIMS, thus it appears to be difficult (or impossible?) to separate doping effects at the poly-insulator interface from quantum-mechanical depletion effects.

4. Direct tunneling gate leakage

Direct tunneling is the most serious quantum effect in MOS devices. A fundamental physical barrier to scaling is given by band-to-band tunneling in the body-to-

drain diode, which limits the maximum body doping and the minimum depletion width. Physics-based modeling of direct and defect-assisted band-to-band-tunneling in pn-junctions (important for GIDL) has been described elsewhere [12]. Gate tunneling leakage may be dominated by resonant processes via defect levels in the oxide [13], if $t_{ox} \geq 3\,\mathrm{nm}$ or, for thinner layers, if the gate bias is low. Direct tunneling determines the off-current (and hence the stand-by power consumption) for SiO_2 thinner than 2 nm in critical regions of the gate. Therefore, alternative materials and stacked dielectrics are considered as future replacements for thermal oxides.

4.1. Analytical models

In the simplest approach, the elastic direct tunnel current through a dielectric layer can be evaluated as (e.g. [14]):

$$j_n = \frac{q m_c^*}{2\pi^2 \hbar^3 \beta} \int dE\, \mathcal{T}(E) \mathcal{F}(E; E_{F,s}, E_{F,g})\,, \tag{7}$$

where $\mathcal{F}(E; E_{F,s}, E_{F,g})$ is a driving force term, which depends on the position of the Fermi levels on the gate and substrate sides (the only non-locality in this model), and $\mathcal{T}(E)$ is the transmission coefficient of the potential barrier produced by the insulating layer. Eq. (7) is for one valley, hence m_c^* can serve as a fit factor. $\mathcal{T}(E)$ has to be expressed by an analytical function, which is only possible for simple potential shapes like trapezoids [15]. In the self-consistent implementation of Ref. [16], Eq. (7) is treated as surface recombination current and solved together with the drift-diffusion equations. In this way the tunnel current is automatically linked to the supply from source/drain (MOSFET) or to the thermal generation in the depletion region of the substrate (MOS capacitor). Model (7) is fast and numerically robust.

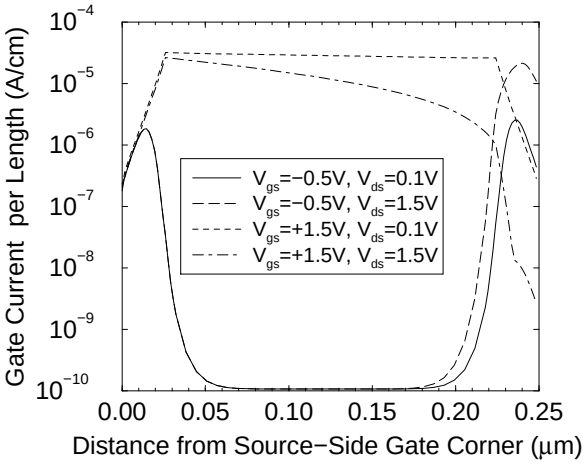

Figure 5. Profiles of gate tunneling current.

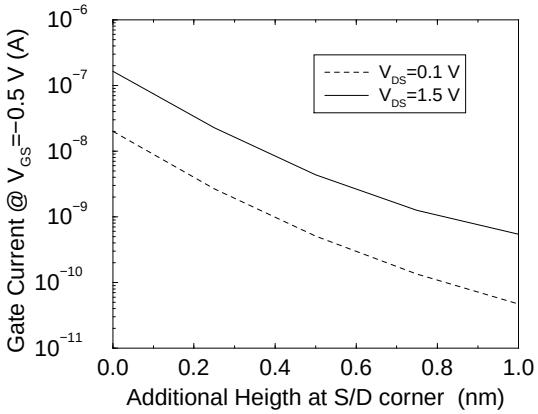

Figure 6. Gate current *vs* increment of t_{ox}.

In MOSFETs, the I_g-V_{gs} characteristics are determined by the balance between in-tunneling at the drain (at high drain bias) and out-tunneling at the source side. A small increase of the gate oxide thickness (e.g. due to poly re-oxidation) in the overlap regions, where the maximum field strength occurs, significantly reduces the direct-tunneling gate current without having a visible influence on the drive current. This is demonstrated in Fig. 6, where a gradually increasing oxide thickness at both gate corners over a distance of 26 nm was assumed

(starting from $t_{ox} = 1.5\,\mathrm{nm}$). The corresponding gate current profiles are shown in Fig. 5. The bell shape is result of the increasing oxide thickness ($\mathrm{t_{ox}} = 2\,\mathrm{nm}$ at the gate corners). An increment of 5 Å over 26 nm drops the leakage current by one order of magnitude.

4.2. Bardeen's method

More complicated potential shapes (e.g. as generated by stacked dielectrics) demand for a numerical solution of the Schrödinger equation. The decay rate from quasi-bound (real) states in the channel to unbound states in the gate can be computed by Bardeen's perturbative method [17] (which is equivalent to ordinary first-order perturbation theory [3]). The tunnel current then reads [13]:

$$j_n = \frac{q\sqrt{2m_g}L}{8\pi m_{ox}^2} \sum_{i,\nu} m_{xy}^{(\nu)} \int dE\, \delta(E, E_i^{(\nu)}) \times \\ \times \left| \Psi_{\tilde{E}} \partial_z \Psi_i^{(\nu)} - \Psi_i^{(\nu)} \partial_z \Psi_{\tilde{E}} \right|^2_{z=z_0} , \qquad (8)$$

where $\delta(E, E_i^{(\nu)})$ contains the difference of the Fermi functions, $\psi_{\tilde{E}}$ is the wave function in the gate ($\tilde{E} = E_i^{(\nu)} + (1 - m_{xy}^{(\nu)}/m_g)E$), and z_0 denotes the interface location.

Fig. 7 shows IV-characteristics of MOS capacitors with different oxide thicknesses obtained with the analytical model (7) and the full quantum-mechanical treatment (8), respectively. A p-substrate with $\langle 100 \rangle$-orientation, $N_A = 10^{18}\,\mathrm{cm}^{-3}$, and $m_{ox} = 0.42\,m_0$ have been used. In Eq. (7) m_c^* had been adjusted once only. Since (8) is solved together with the drift-diffusion equations, emptied states $\Psi_i^{(\nu)}$ are refilled by SRH generation, which leads to the behavior of a reversed-biased pn-diode at positive voltages in Fig. 7. Self-consistency is at the expense of numerical robustness; the same items as discussed in Subsection 2.2 apply

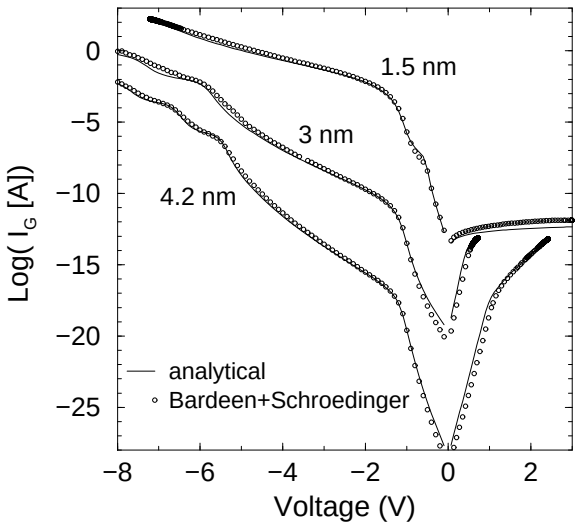

Figure 7. Direct tunnel currents of MOS capacitors.

to model (8). Moreover, at strong negative gate bias the necessary number of eigensolutions can be very large (up to about 300 for the curves presented in Fig. 7), because most of the electrons injected from the gate arrive on the silicon side with a high energy.

4.3. Gamow's method

An alternative approach is based on Gamow's picture of nuclear decay [18]. The electron tunnel current density is calculated as [19]

$$j_n = \frac{q}{\pi\hbar^2\beta} \sum_{i,\nu} \frac{m_{xy}^{(\nu)}}{\tau_i^{(\nu)}} \mathcal{F}\left(E_i^{(\nu)}; E_{F,s}, E_{F,g}\right) , \qquad (9)$$

where symbols have the same meaning as before. The new quantity is the resonance lifetime $\tau_i^{(\nu)} = \hbar/2\Gamma_i^{(\nu)}$ of a quasi-bound state in the channel leaking into the gate. The resonance width $\Gamma_i^{(\nu)}$ is found numerically by solving the Schrödinger equation in a domain that covers substrate, insulator, and gate. Eq. (9) relies on the assumption that the resonance widths of the quasi-bound states are much smaller than their en-

ergies: $\Gamma_i^{(\nu)} \ll E_i^{(\nu)}$, which restricts the method to the range of deep inversion. In practice, the case of in-tunneling from the gate (near the drain, see Fig. 5) is more relevant. This, and the numerical expense to trace the resonance peaks and spectral widths, make the method less suited for integration into TCAD packages. A careful comparison between Bardeen's method and the resonance method in a range, where the latter was meaningful, yielded a slight disagreement only ($\approx$ factor 2-3, see Fig. 8). For this comparison, identical MOS structures and parameters were used, as well as a data interface between the simulator [4] and the program described in Ref. [19] in order to ensure identical potential profiles. For

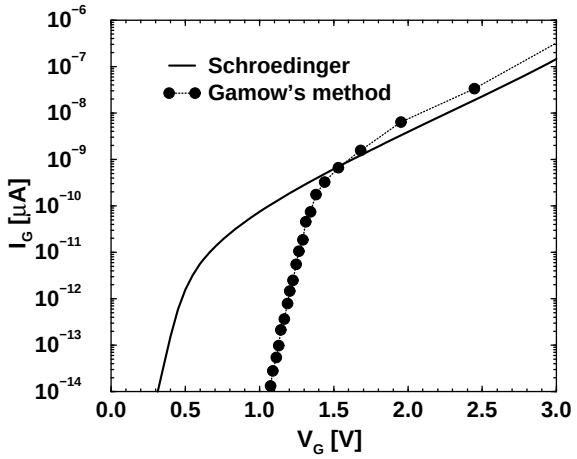

Figure 8. Comparison between 'Schrödinger' and 'Gamow'.

Fig. 8 a capacitor with $N_A = 10^{15}\,\mathrm{cm}^{-3}$, A$= 2 \times 10^5 \mu\mathrm{m}^2$, $t_{ox} = 3\,$nm was used.

4.4. Density gradient method

The DG method describes tunneling by a normal drift-diffusion current. This current flows through the barriers, which are strongly reduced by the quantum potential Λ (see Fig. 9). If tunneling is not described as surface recombination (as done in [20]), the current equation has to be solved

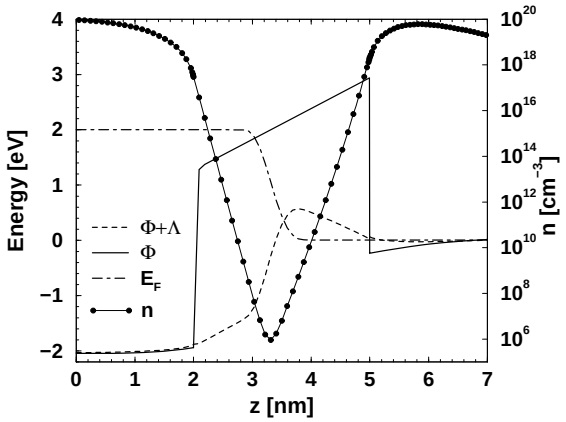

Figure 9. Density and potential profiles.

in the barriers, too, which requires that the oxide is treated as a wide band gap semiconductor. Quasi-Fermi levels are then also defined in the barriers and they drop according to the local 'tunnel' current. Fig. 9 illustrates the potential and density profiles across an NMOSFET with 3 nm oxide thickness at 2 V gate voltage [21, 4]. It is striking that the density decays into the barrier from both sides, although the tunnel current flows from channel to gate. Since the 'tunnel' current is determined by the lowest value of the density in the barrier, the tunneling length appears to be roughly one half of the actual tunneling length. This point has been addressed in the literature [20] by introducing different carrier types according to their tunneling direction. Such an extension of the DG model is bound to 1D problems, since the concept of 'forward' and 'backward' has no obvious generalization to multi-dimensional device simulation.

Fig. 10 shows that the DG method is able to reproduce gate leakage currents in MOSFETs. NMOSFETs with a gate length of 300 nm, a gate width of 1 μm and oxide thicknesses of 2 nm and 3 nm, respectively, were simulated with both the self-consistent Schrödinger method and the DG model. To fit the latter to the former, the mobility in the

oxide was adjusted to $\mu_{\rm n} = 0.05\,{\rm cm}^2/{\rm Vs}$ using the case $t_{\rm ox}$ =2 nm [21, 4]. The agreement is within one order of magnitude.

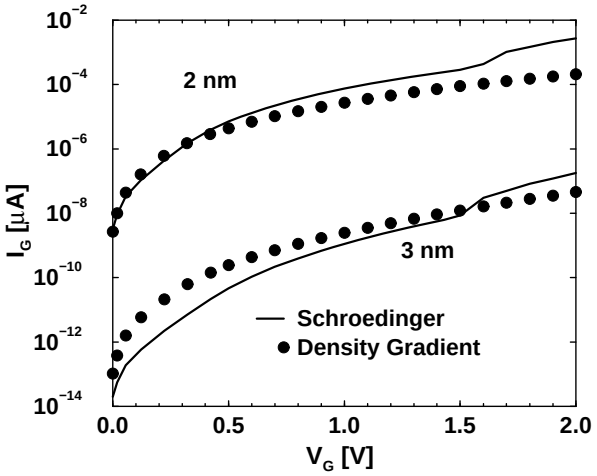

Figure 10. Gate tunneling currents in NMOSFETs.

The predictibility of gate currents is certainly worse than one order of magnitude, despite the similar results obtained by rather different methods. In reality, insulating layers are never defect-free and parameters (like $m_{\rm ox}$, $\epsilon_{\rm ox}$, $E_{g,{\rm ox}}$) proven for bulk, become uncertain in the case of $t_{ox} \approx 10$Å.

5. Acknowledgements

Collaboration with Dr. Andreas Wettstein (ISE AG Zurich), Dr. Eugeny Lyumkis, and Dr. Oleg Penzin (ISE Corp. San Jose) benefited the paper substantially.

[1] M. J. VAN DORT, P. H. WOERLEE, and A. J. WALKER. *A simple model for quantisation effects in heavily-doped silicon MOSFETs at inversion conditions.* Solid-State Electronics **37**(3):411–414 (1994).

[2] F. STERN. *Self-consistent results for n-type Si inversion layers.* Phys. Rev. B **5**(12):4891–4899 (1972).

[3] A. WETTSTEIN. *Quantum Effects in MOS Devices.* Ph.D. thesis, ETH Zürich (2000). Hartung-Gorre, Konstanz. ISBN 3-89649-566-6.

[4] ISE Integrated Systems Engineering AG. *DESSIS 7.0 reference manual* (2001).

[5] C. L. GARDNER and C. A. RINGHOFER. *Smooth quantum potential for the hydrodynamic model.* Phys. Rev. E **53**(1):157–167 (1996).

[6] M. G. ANCONA and G. J. IAFRATE. *Quantum correction to the equation of state of an electron gas in a semiconductor.* Phys. Rev. B **39**(13):9536–9540 (1989).

[7] J.-R. ZHOU and D. K. FERRY. *Ballistic phenomena in GaAs MESFETs: Modelling with quantum moment equations.* Semicond. Sci. Technol. **7**(3B):B546–B548 (1992).

[8] A. WETTSTEIN, A. SCHENK, and W. FICHTNER. *Quantum Device-Simulation with the Density-Gradient Model on Unstructured Grids.* IEEE Trans. Electron Devices **48**(2):279–284 (2000).

[9] A. S. SPINELLI, A. PACELLI, and A. L. LACAITA. *Polysilicon Quantization Effects on the electrical properties of MOS transistors.* IEEE Trans. Electron Devices **47**(12):2366–2371 (2000).

[10] M. G. ANCONA, Z. YU, W.-C. LEE, R. W. DUTTON, and P. V. VOORDE. *Simulation of quantum confinement effects in ultra-thin-oxide MOS structures.* J. of Technology Comp. Aided Design (11) (1999). See http://www.ieee.org/products/online/journal/tcad/.

[11] A. WETTSTEIN, A. SCHENK, A. SCHOLZE, G. GARRETÓN, and W. FICHTNER. *Charge carrier quantization effects in double-gated SOI MOSFETs.* In H. Z. Massoud, H. Iwai, C. Claeys, and R. B. Fair, eds., *ULSI Science and Technology 1997*, pp. 613–621. The Electrochemical Society, Inc., Pennington, NJ, USA (1997).

[12] A. SCHENK. *Advanced Physical Models for Silicon Device Simulation. (Computational Microelectronics, ed. by S. Selberherr)*, Springer Wien New York (1998).

[13] A. WETTSTEIN, A. SCHENK, A. SCHOLZE, and W. FICHTNER. *The influence of localized states on gate tunnel currents – modeling and simulation.* In *SISPAD 1997 Technical Digest*, pp. 101–104.

[14] P. J. PRICE and J. M. RADCLIFFE. IBM Journal Oct, (1959).

[15] K. H. GUNDLACH. Solid-State Electronics **9**:949 (1966).

[16] A. SCHENK and G. HEISER. *Modeling and simulation of tunneling through ultra-thin gate dielectrics.* J. Appl. Phys. **81**(12):7900–7908 (1997).

[17] J. BARDEEN. Tunneling from a Many-Particle Point of View. *Phys. Rev. Lett.*, **6**(2):57–62 (1961).

[18] G. A. GAMOW. Zs. Phys. **51**(3-4):204 (1928).

[19] W. MAGNUS and W. SCHOENMAKER. Full quantum mechanical model for the charge distribution and the leakage currents in ultrathin metal-insulator-semiconductor capacitors. *J. Appl. Phys.*, **88**(10):5833–5842 (2000).

[20] M. G. ANCONA, Z. YU, R. W. DUTTON, P. J. V. VOORDE, M. CAO, and D. VOOK. *Density-gradient analysis of MOS tunneling.* IEEE Trans. Electron Devices **47**(12):2310–2319 (2000).

[21] A. WETTSTEIN, O. PENZIN, and E. LYUMKIS. *Integration of the Density Gradient Model into a General Purpose Device Simulator.* Submitted to VLSI Design