Eur. Phys. J. AP **15**, 117–121 (2001)

THE EUROPEAN PHYSICAL JOURNAL APPLIED PHYSICS

Simulation of 2D quantum effects in ultra-short channel MOSFETs by a finite element method*

A. Poncet[a], C. Faugeras, and M. Mouis

Laboratoire de Physique de la Matière[b], INSA-Lyon, 20 avenue Albert Einstein, 69621 Villeurbanne, France

Received: 20 November 2000 / Revised: 18 April 2001 / Accepted: 27 April 2001

Abstract. This paper presents a flexible numerical technique which is especially suited to analyze lateral modulation of quantum effects in short channel MOS transistors. We discuss boundary conditions for the Schrödinger equation and the impact of the finite element meshing. We show how channel length shortening alters the sub-band structure, thus giving an evaluation of the limits of a 1D quantum approach.

PACS. 85.30.De Semiconductor-device characterization, design, and modeling – 02.70.Dh Finite-element and Galerkin methods

1 Introduction

It is now recognized that quantization must be accounted for to analyze correctly the operation of field-effects transistors using inversion channels along an insulating or wide bandgap semi-conducting layer (MOSFETs and MODFETs). It has been shown for instance that the gate insulator thickness of a MOSFET, as deduced from capacitance measurements, was significantly over-estimated if quantization was neglected. Although quantization in MOS channel was present from the origin (and usually neglected), the corrections which it introduces to standard extraction procedures are no longer negligible, due to the reduction of gate oxide thickness in advanced technologies.

Different simulation approaches have been used in the literature. Due to carrier confinement against the gate insulator, quantum effects are usually simulated using self consistent numerical solutions of Poisson and Schrödinger equations in 1D across the channel [1,2]. This approach may not remain valid in very short gate length MOSFETs, where significant confining field variations are observed between source and drain, on very short distance scales. Besides, a strict 1D approach make it impossible to account for effects such as lateral variations of oxide thickness or dopant profiles.

However, only few attempts have been made to account for the lateral variation of quantum confinement along the channel [3].

This paper presents results of a full-2D simulation of quantum effects and of the influence of a gate length reduction down to a few ten nanometer. This work differs from previous attempts to solve Schrödinger equations in two dimensions [3,4] by the use of finite elements which are especially well suited to analyze devices of arbitrary shape, such as self- organized quantum boxes or MOSFETs with non uniform gate oxide.

2 Numerical methods

2.1 Galerkin formulation of Schrödinger equation

Let us consider a domain Ω (of any dimensionality) in which quantum effects need to be simulated. The Green formula can be used to integrate Schrödinger equation over Ω with a test function w in a suitable functional space [7]. The resulting Galerkin formulation of Schrödinger equation for any valley in the silicon lattice reads then:

$$\int_\Omega \left(-\frac{\eta}{2} \sum_{i=1}^{n} \frac{1}{m^\star_{xi}} \frac{\partial \psi}{\partial xi} \frac{\partial w}{\partial xi} + (E - V) \cdot \psi . w \right) \cdot \partial\Omega + \oint_\Gamma \frac{\eta}{2m^\star_\eta} \frac{\partial \psi}{\partial \eta} \cdot w \cdot \partial\Gamma = 0 \quad (1)$$

where Γ denotes boundaries and interfaces of Ω, and η is the outer normal direction to this boundary. We accounted for the ellipsoidal symmetry of the conduction band structure: $m^\star_{xi}$ is the electron effective mass in direction x_i for the valley under consideration. V is the potential energy deduced from Poisson's equation, and E is the eigen energy associated with wave function ψ.

We used classical boundary conditions for Schrödinger equation:

- $\psi = 0$ for an infinite barrier; (2)
- continuity of $\frac{1}{m^\star_\eta} \frac{\partial \psi}{\partial \eta}$ across interfaces; (3)
- $\frac{\partial \psi}{\partial \eta} = 0$ for reflecting boundaries. (4)

* This paper has been presented at 3[es] Journées Nationales "Hétérostructures à semiconducteurs IV-IV", Orsay, July 2000.

[a] e-mail: poncet@lpm.insa-lyon.fr

[b] UMR-CNRS 5511

All three conditions lead the boundary integrals in (1) to vanish, making implementation in a finite element simulator straightforward.

Condition (4) is especially well suited to account for symmetries. It can be used to save CPU time by reducing the size of the simulation domain to only half a device (in the case of symmetric devices such as double-gate MOSFETs or Permeable Base Transistors, ...) or by shortening ohmic areas.

If needed, periodic boundary conditions might also be introduced to simulate Bloch waves. However, with such boundary conditions, the matrices loose their band structure. This makes implementation more complex and more CPU time consuming.

2.2 Finite element simulation

The finite element discretization of equations (1) to (4) is straightforward: a basis of piecewise polynomial shape functions w_i is constructed. These functions are piecewise linear on a triangular or tetrahedral mesh, and piecewise bi-linear on a quadrangular or hexahedral mesh. Wave functions are projected on this basis to get a discrete eigensystem. However, attention must be paid to the mapping from eigenvectors to wave functions. In this linear algebraic problem, the eigenvectors do not provide directly the wave function values at mesh points. If numerical integration uses a trapezoidal rule, the matrix of $\int_\Omega w_i w_j \partial\Omega$ terms is diagonal, but is not the identity matrix. The component number i of any eigenvector is the product of the corresponding wave function at node number i by a factor $\int_\Omega w_i^2 \partial\Omega$. On the other hand, if the integrals are computed exactly, the above mentioned matrix is even not diagonal, and must be first diagonalized. This is the only CPU time overhead in comparison with finite difference schemes.

2.3 Discussion of alternative methods

An alternative way to discretize Schrödinger equation has been proposed in the literature [3,4]. There, the piecewise polynomial shape functions w_i are replaced by sine functions, with space coordinates splitting, in order to decorrelate the minimum wave function wavelength from the choice of the mesh size. However, although attractive at first sight, this method gives full matrices while finite element and finite difference schemes lead to sparse matrices. Moreover, the only boundary which can be handled is condition (2), and it seems very difficult to extend this method to arbitrarily shaped structures. Finally, the smallest period which can be introduced with this method is in fact the same as with our more classical discretization scheme. This can be illustrated in the 1D case. If L is the structure length, the smallest wavelength which can be simulated is $2L/N$ in both cases, N representing the number of terms in the sine expansion in one case and $N+1$ representing the number of mesh points in our case. It means that both methods actually lead to similar matrix sizes, so that the only potential drawback of the finite element method is in fact minor in comparison with its advantages.

In another respect, a finite difference method could be even simpler and slightly faster (as explained in the previous paragraph). We did not choose this approach however, in order to keep the latitude of simulating arbitrary shaped devices.

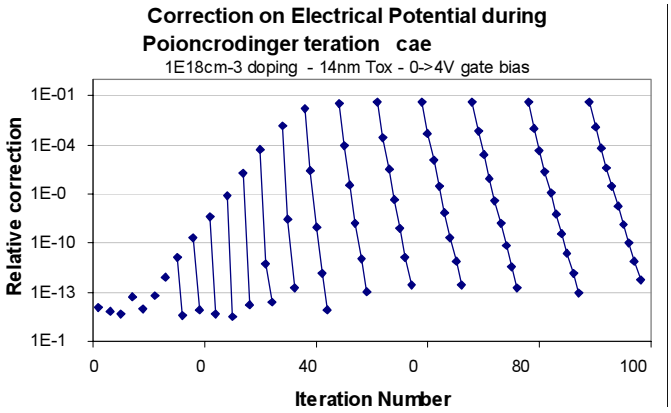

Fig. 1. Convergence of the Schrödinger-Poisson decoupling scheme for 1D simulation of a NMOS capacitor Q-V characteristic.

2.4 Iterative decoupling scheme

Several methods have been proposed to solve the coupled system of Poisson and Schrödinger equations iteratively. The most straightforward method consists in re-injecting the carrier concentrations calculated from Schrödinger equation in the right hand side member of Poisson equation. However, this is known to result in instabilities which can be overcome by using under-relaxation methods. Convergence is then rather slow. A much more efficient method has been proposed in [5]. It consists in developing a Newton-like method in which the Jacobian matrix is replaced by a first order approximation. A huge amount of mathematical work is needed to end up with approximations that are both valid and computable.

Here, we propose a new method which is altogether simple and fast. Once electron concentration has been computed by summing up all the wave functions contributions, a "Fermi-like" potential is extracted using the relationship which relates carrier concentration to the electrostatic potential when a volume Boltzman statistics holds. This quantity (which is physically meaningless in the present situation, and must only be seen as a variable mapping) is injected in Poisson equation, as usually done in classical device simulators. This make Poisson equation highly non-linear but very easy to solve with a pure Newton-Raphson procedure. The procedure is iterated until convergence is reached.

Convergence is extremely fast, as shown in Figures 1 and 2. Figure 1 shows the number of iterations needed to bring the relative correction below 10^{-12} during the simulation of the Q-V characteristic of an NMOS capacitor

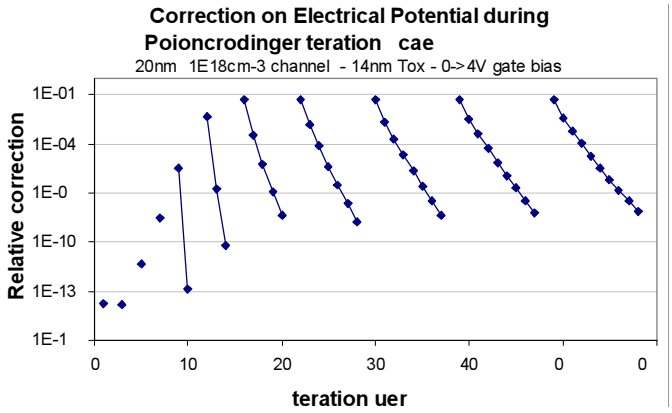

Fig. 2. Convergence of the Schrödinger-Poisson decoupling scheme for the 2D simulation of a short channel NMOS transistor gate characteristic.

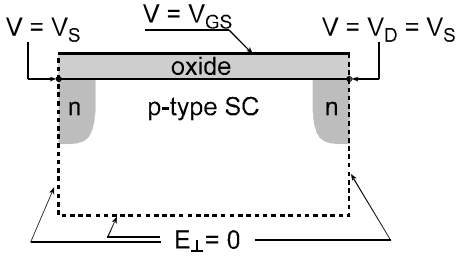

Fig. 3. Boundary conditions for Poisson equation. For Schrödinger equation, we used Neuman boundary conditions on the outer boundaries.

(with 21 successive voltage steps). Dimensions and doping levels were taken in [6] so that we could check our results. The oxide layer was 14 nm thick, so that high voltages (up to 4 V) needed to be applied to the gate. In the weak inversion regime, only a couple of iterations are required. The error exhibits an exponential decay even at high gate voltages. We obtained a very similar behavior in 2D, during the simulation of the gate characteristics a short channel NMOS transistor, with 11 successive voltage steps (Fig. 2). This proves that this method is efficient in both the 1D and 2D cases.

3 Simulation results

When narrow channel effects are neglected, it becomes obviously useless to solve the problem in 3D. The 2D Schrödinger equation is known to be appropriate to study quantum wire structures, *i.e.* carrier confinement in 2 directions. The use of this 2D approach to investigate quantum effects in a short channel MOSFET is less straightforward, because carriers are not confined laterally in the channel. Therefore, boundary conditions and meshing must be defined carefully.

The boundary conditions used for Poisson and Schrödinger solutions are given in Figure 3. These boundary conditions will be discussed below with the support of simulation results.

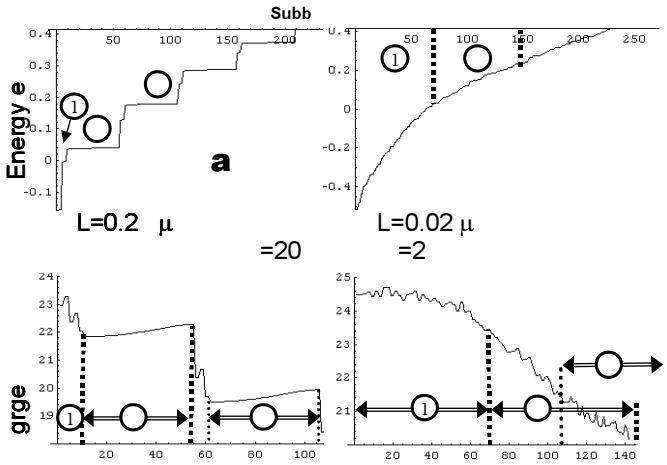

Fig. 4. Eigen energy values as a function of the sub-band index (a, b) and contribution to the total electron charge of the corresponding wave function (c, d) for long and short channel length (0.92 μm for a-b and 20 nm for b-d).

Here, we restrict ourselves to equilibrium conditions ($V_{\rm DS}$ = 0 V). Our method can be used then to calculate the 2D distribution of electrons and holes, together with the corresponding 2D potential and electric field distributions, as a function of gate voltage. This gives for instance indications about gate control of the channel carrier charge. Electrical parameters such as the sub-threshold slope and threshold voltage at low $V_{\rm DS}$ can be extracted and short-channel effects at low $V_{\rm DS}$ can be evaluated: indeed, 2D effects, resulting in a parasitic control of the source/channel potential barrier are accounted for. Moreover, the existence of non-zero wave- functions in regions of the device where the sub- band energy is in the bandgap, can bring an evidence of the presence of quantum coupling between some regions of the devices and of possible tunneling effects. However, full current-voltage characteristics will be available only once this model has been coupled to a quantum transport model. The following discussion focuses on the numerical conditions to get a physically significant solution under thermodynamic equilibrium conditions, with given computer resources (limited number of mesh points).

In order to investigate how the calculated eigen energy spectrum depends on channel length, we performed 2D simulations with electrical channel lengths (L) ranging from 20 nm to 1 μm, with a fixed number of mesh points in the direction of the channel (∼50). Channel doping was 10^{17} cm^{-3}, and the gate oxide was 2.5 nm thick. The NMOS transistor was biased with 1 V on the gate (assuming zero work function and no interface charges). This bias voltage corresponds to inversion conditions.

3.1 Long channel

The 2D solution provides a series of sub-bands with a minimum energy ε_i (corresponding each to a given wave function Ψ_i in the (x,y) plane) and a continuum for wave

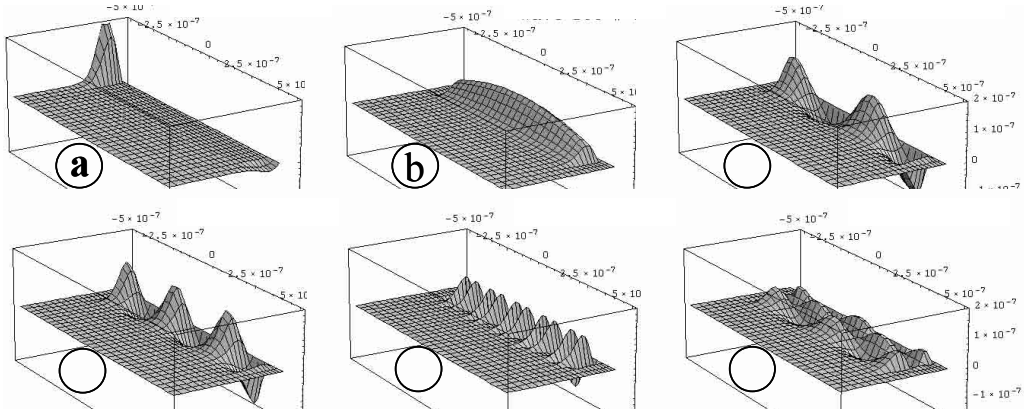

Fig. 5. Example of wave functions obtained (a) in the S/D ohmic regions (region 1 of the energy spectrum in Fig. 4), (b)-(e) in the first sub-band subset (wave functions with one lobe perpendicular to the gate, region 2 of the energy spectrum) and (f) in the second sub-band sub-set (wave functions with two lobes perpendicular to the gate, region 3 of the energy spectrum).

vector k_z in the z direction (along the device width). Figure 4a shows the ε_i values obtained as a function of the sub-band index in the long channel case (1 μm). The energy spectrum is clearly stepwise.

For moderate gate voltages, the lowest energies (region 1 of the energy spectrum) correspond to wave functions which vanish outside the source and drain areas (such as the one shown in Fig. 5a). Next come energies which are associated to wave functions with only one lobe in the direction normal to Si/SiO_2 interface (such as those shown in Figs. 5b to 5e). Then come energies corresponding to wave functions that show two transverse lobes (such as in Fig. 5f), and so on...

Therefore, the band structure degenerates into a series of sub-band sub-sets. Each sub-set corresponds to the sub-band which would be computed using a 1D approach. However, each sub-set contains a finite number of sub-bands, which is equal to the number of mesh points in the channel direction. Moreover, each wave function in this sub-set has a similar weight in the total carrier concentration (Fig. 4c). In the case shown, the 2D numerical solution is physically meaningless, due to mesh-induced truncation at the top of each sub-set. Indeed, as explained below, the criterion to get physically significant results is the mesh size. Therefore, simulation of long channel MOSFETs (with local non-uniformities for instance) is very demanding in computer resources. With 100 nodes, strong errors are still observed at 120 nm channel length.

3.2 Short channel

The picture is rather different for short channel MOSFETs. The first visible results is that the wave functions become strongly two- dimensional, as shown in Figure 6.

However, it should be noted that, as in the long channel case, the wave functions associated to S/D ohmic regions vanish in the channel while channel wave function vanish in the source and drain. The fact that we chose zero normal derivative boundary conditions at the source and drain outer boundary (instead of periodic boundary conditions) is therefore of little importance: these boundary conditions do not disturb results in the channel (unless the ohmic region size is really too small). However, the larger these ohmic regions, the more eigen values will be associated to them. It must therefore be kept in mind that if they are taken too large, it will be necessary to calculate a huge number of eigen energy before a sufficient number of channel eigen energies has been obtained to get physically significant results in the channel.

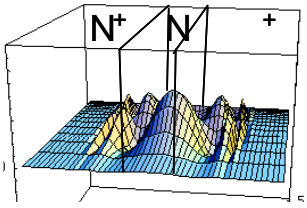

Fig. 6. An example of a 2D wave function calculated in the channel of a MOSFET (with an effective channel length of 20 nm).

The other important result is that the eigen energy stepwise variation is progressively smoothed (Fig. 4b) and that the energy spectra of the different sub-sets overlap: as shown in Figure 4d, the sub-set corresponding to wave functions with two lobes perpendicular to the oxide (region 3) starts before the sub-set corresponding to wave functions with only one lobe (region 2) has ended.

As explained before, this 2D approach cannot be used for very long gates, due to the mesh-induced truncation of the energy spectrum of each sub-set. Let us now discuss in more details the conditions which must be fulfilled to get physically significant results in the channel. For this purpose, we consider the respective contributions to electron concentration of the first and last computed sub-bands in sub-set S_i, which we name here $\alpha_{\text{low}}(S_i)$ and $\alpha_{\text{high}}(S_i)$. The ratio $\alpha_{\text{high}}(S_i)/\alpha_{\text{low}}(S_i)$ reflects the error due to mesh induced truncation in S_i. Figure 7 shows the variation of $\alpha_{\text{high}}(S_1)/\alpha_{\text{low}}(S_1)$ with mesh size for different numbers

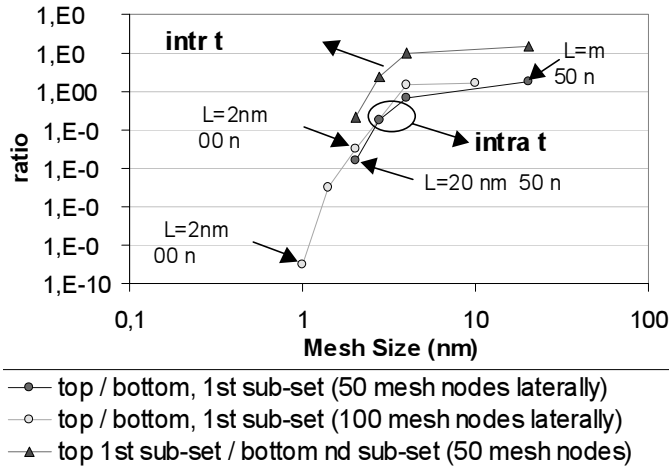

Fig. 7. Evaluation of the mesh truncation error to the calculation of the electron charge as a function of mesh size, channel length and number of nodes.

of mesh nodes (50 and 100) in the lateral direction. This ratio depends on mesh size rather than on gate length or mesh number. This was expected from first order evaluation. Indeed, the minimum simulated wavelength is directly correlated to mesh size and the highest energy wave function in a given subset is expected to have the shortest "wavelength" in the lateral direction. It is therefore natural to get a correlation between the residual error in a given subset and the lateral mesh size. However, speaking of wavelengths here is an image as the 2D functions are not really periodical and it is interesting to note that this trend remains valid in strongly 2D situations. Given a number of nodes, the diagram of Figure 7 provides the maximum channel length which can be simulated with a given accuracy on electron distribution.

In addition, note that if several sub-set are occupied, the condition $\alpha_{\text{high}}(S_1)/\alpha_{\text{low}}(S_2) \ll 1$ indicates that all significant sub-bands have actually been computed, and gives a validity criterium for the whole computation.

4 Conclusions

It can be concluded that 2D numerical solution of the coupled Poisson and Schrödinger equations may be obtained with reasonable computing resources (here, simulations of the NMOS structure were done on a desktop computer). Provided a sufficiently small mesh size has been chosen, physically significant solutions can be obtained for nanoscale gate lengths, in the range where a 1D approach is actually inadequate.

References

1. M. Moglestue, J. Appl. Phys. **59**, 3175 (1986).
2. C. Raynaud, J.-L. Autran, P. Masson, M. Bidaud, A. Poncet, Mat. Res. Soc. Symp. Proc., Vol. 592, Boston, 1999.
3. A. Abramo, A. Cardin, L. Selmi, E. Sangiorgi, IEEE Trans. Electron Devices **47**, 1858 (2000).
4. A. Abou-Elnour, K. Schuenemann, J. Appl. Phys. **74**, 3273 (1993).
5. A. Trellakis, A.T. Galick, A. Pacelli, U. Ravaioli. J. Appl. Phys. **81**, 7880 (1997).
6. M.J. Van Dort, P.H. Woerlee, A.J. Walker, Solid-State Electron. **37**, 411 (1994).
7. P. Ciarlet, The finite element method for elliptic problems, *Studies in Mathematics and its applications* (North Holland, 1978), Vol. 4.