THREE-DIMENSIONAL, SELF-CONSISTENT MODELING OF MOS DEVICES UNDER NON-ISOTHERMAL REGIME*

Paolo Ciampolini, Anna Pierantoni, Giorgio Baccarani

*This work was supported by CNR, under the “P. F. MADESS” and by ST Microelectronics S.p.A.

1. Introduction

Due to the dissipated power density, self-heating effects occurring in densely-packed VLSI integrated circuits deserve careful consideration. Numerical simulation of thermo-electric features in semiconductor devices have been presented in the past [1-4]: however, the inherently three-dimensional nature of the heat diffusion very often makes the commonly-adopted 2D approximation too crude. In this paper, a 3D device simulator is used to investigate such effects: to this purpose, the Drift-Diffusion transport model has been supplemented with the thermal-diffusion equation in a self-consistent way. The physical model and the relative assumptions are discussed in section 2, while section 3 deals with discretization issues. The electro-thermal simulation of an nMOS large-current driver is illustrated in section 4.

2. The physical model

In order to account for the non-isothermal condition, the classical Drift-Diffusion (DD) transport model has to be modified as follows:

\[
\text{div} (\text{grad} \varphi) = -\frac{q}{\epsilon_s} (p - n + N_D^+ - N_A^-)
\]

\[
\text{div}(J_n/q) = R - G
\]

\[
\text{div}(\kappa(T)\text{grad} \, T) = -P
\]

where:

\[
J_n = -q\mu_n \, n \, \text{grad} \, \varphi + qD_n \, n \, \text{grad} \, n + qnD_T \, \text{grad} \, T
\]
For the sake of conciseness, the above equations have been specialized to the steady-state, single-carrier case; the additional terms dealing with heat diffusion are enclosed in frames. The symbols have their usual meanings: $\phi$ is the electric potential and $T$ the lattice temperature; equation (1) is the Poisson equation, while equation (2) expresses the electron current continuity, $R - G$ is the net recombination rate and $J_n$ the electron current density, defined as in (4). The extra term appearing in the latter accounts for the so-called thermoelectric effect [5], $D_T$ is the thermal diffusion coefficient. In the heat equation (3), $\kappa(T)$ represents the thermal conductivity, while the driving force $P$ represents the heat generation rate: many authors suggested different expressions for $P$, which have been recently reviewed by Wachutka [6] in the framework of a rigorous thermodynamic approach. For MOS devices, a few remarks can be made:

- the net recombination rate is assumed to be negligible
- the electron mobility is described by means of the CVT model [7]
- the term $P$ accounts for the Joule heat only, i.e.:

$$P = E \cdot J_n$$  \hspace{1cm} (5)

Such expression agrees with most of the above-referenced works in the steady-state, homogeneous ($R - G = 0$) case which will be treated in what follows.

Assuming [8]:

$$D_T = \frac{1}{T} D_n$$

equation (4) can be rewritten as follows:

$$J_n = q\mu n \left[ -\text{grad} \phi + \frac{k_B T}{q} \left( \frac{1}{n} \text{grad} n + \frac{1}{T} \text{grad} T \right) \right]$$  \hspace{1cm} (6)

Equation (6) allows to estimate the magnitude of the contribution to $J_n$ coming from the thermoelectric field, by comparing it with the standard Drift-Diffusion terms; Fig. 1 shows the ratio between the two driving forces:

$$Z = \frac{\frac{1}{T} \text{grad} n}{\frac{1}{T} \text{grad} T}$$  \hspace{1cm} (7)

computed along the channel of a 1$\mu$m MOSFET: the electron concentration profile is also shown for reference. It is worth remarking that inside the transistor active region, the diffusive term is at least 50 times greater than the thermoelectric one, so that the total current density is affected by the latter to a fairly small extent. Inside the drain and source diffusions, where such a ratio decreases significantly, transport is largely dominated by the ohmic component.

The lattice temperature affects most coefficients of equations (1), (2) and (4): the heat equation is therefore solved following a decoupled strategy. First, the system of equations (1), (2) and (4) is solved by keeping $T$ constant; then equation (3) is solved to update the temperature and the procedure is iterated until the unknowns ($\phi$, $n$, $T$) converge to the solution.
3. Discretization

All equations are discretized according to the so-called Box Integration Method (BIM, [9]), where 3D boxes are built-up starting from prismatic elements [10].

To this purpose, it is worth observing that the above equation set can also be regarded as derived from the so-called "Hydrodynamic" model described, e.g., in [11], by assuming local thermal equilibrium inside the device (i.e., the carrier temperature equals the lattice one) and neglecting some less-significant terms, so that a similar discretization strategy can be applied. By expressing the term \( P \) as:

\[
E \cdot J_n = -\text{div}(\varphi J_n) + \varphi \text{div}J_n
\]  

and again neglecting generation and recombination (yielding \( \text{div}J_n = 0 \)), equation (5) can be rewritten as follows:

\[
\text{div}(\kappa(T)\text{grad}T) = \text{div}(\varphi J_n)
\]  

Integrating (9) over each element \( \Omega_i \), and applying the divergence theorem, one finds:

\[
\int_{\Gamma_i} \kappa(T)\text{grad}T \cdot i_n \, d\Gamma_i = \int_{\Gamma_i} \varphi J_n \cdot i_n \, d\Gamma_i
\]  

where \( i_n \) is a normal versor, outward-oriented with respect to the external surface \( \Gamma_i \). Equation (10) is discretized by assuming a piecewise bi-linear approximation for \( T \) inside each prism. The discrete form of (10) then reads:

\[
\sum_{j \neq i} < \kappa_{ij} > \frac{a_{ij}}{s_{ij}} (T_j - T_i) = \sum_{j \neq i} [< \varphi_{ij} > a_{ij} J_{ij}]
\]  

where \( < \kappa_{ij} >, < \varphi_{ij} > \) are the average on side \( s_{ij} \), and \( J_{ij} \) the related current-density projection. In (11) the sums are extended to the nearest neighbours of node \( i \) and \( a_{ij} \) represents the cross section associated with side \((i,j)\).

By normalizing the nodal potential:

\[
u_i = \frac{q\varphi}{kB}T_0
\]

and the nodal temperature:

\[
r_i = \frac{T_i}{T_0}
\]

with respect to the room temperature \( T_0 \), and introducing the auxiliary quantities:

\[
u_{ij} = u_j - u_i
\]
\[r_{ij} = r_j - r_i
\]
\[\Theta_{ij} = \frac{\log(r_j/r_i)}{r_{ij}}
\]
\[\Delta_{ij} = \Theta_{ij}(u_{ij} - 2r_{ij})
\]

the current density \( J_{ij} \) can be expressed as follows [12]:

\[
J_{ij} = \frac{qD_n}{\Theta_{ij}s_{ij}} \left[ B(\Delta_{ij}) \frac{n_j}{r_j} - B(\Delta_{ji}) \frac{n_i}{r_i} \right]
\]
where:

\[ B(\Delta_{ij}) = \frac{\Delta_{ij}}{\exp(\Delta_{ij}) - 1} \]

is the Bernoulli function. Usually only a thin layer of the silicon wafer is accounted for in the device simulation: in order to account for the heat flowing toward the bulk (or any other heat sink) boundary conditions are imposed which account for the external thermal resistance per unit area \( r_{th} \):

\[ \kappa(T) \text{grad}(T) \cdot i_n = \frac{T - T_o}{r_{th}} \]

Equation (11) originates a set of algebraic linear equations, whose coefficient matrix turns out to be symmetric and positive definite. The ICCG method [13], already employed for the solution of the Poisson equation, lends itself to the solution of such a linear system as well.

**Results**

The above technique has been applied to the simulation of a MOSFET whose layout, shown in Fig. 2, has been conceived as a part of an IC output buffer fabricated with a 1\( \mu \)m nMOS technology. The fingered structure allows a large current driving capability within a limited chip area, which leads to a fairly large power density.

In our example, only the elementary cell enclosed between the dashed lines has been simulated: in order to account for the thermal coupling among adjacent cells, Neumann-type boundary conditions have been imposed to the heat fluxes across the dashed sections. I.e., no heat is exchanged between neighboring cells, which is equivalent to assuming that the pattern of the fingers is indefinitely replicated. Thermal resistances have been accounted for at the remaining edges of the cell: namely, the \( r_{th} \) associated with the edges facing the substrate has been obtained by assuming a substrate thickness of 500 \( \mu \)m, while much higher values have been imposed at the upper, passivated layers. Finally, low-valued thermal resistances have been connected to the drain and source electrodes.

Fig. 3 shows the 13108-node, prismatic-element mesh: computations required roughly 1 CPU hour per bias point (on a IBM RISC6000/320 Workstation). The overhead induced by the heat equation, with respect to the DD solution, is fairly moderate: an average of two additional DD iterations (usually much more quickly converging than the first one) was required by the algorithm.

Fig. 4 shows the computed temperature distribution: power dissipated mainly at the drain edges along the channel heats the device up to some 350°K \( (V_{GS} = V_{DS} = 5\text{V}) \). The temperature decreases smoothly toward the bulk, while the cooling effect of the drain and source metallization is fairly evident, as well as some corner effects at the channel twists: at the convex drain corner, where the current and field lines concentrate, the temperature rises above the channel average value, which results in the "hot spot" showing up at the lower right corner of the channel. On the contrary, current lines diverge where the drain edge is concave, thus reducing the current density and, in turn, the temperature.

Temperature ranges inside the simulated devices are shown in Fig. 5; Fig. 6 shows the computed currents, which are compared with DD-computed ones: mainly due to the mobility degradation, the electro-thermal simulation provides significantly lower currents, the maximum discrepancy being in the order of 15%.

Finally, Fig. 7 shows the temperature distribution at the Si – SiO\(_2\) interface: the plot has been obtained by mirroring three times the elementary cell, and allows for a more comprehensive...
insight of the thermal phenomena: two “fingers” are visible, depicted from the temperature ridges along the channel regions. The foremost finger belongs to the drain region and its tip exhibits two sharp temperature peaks, while the rear finger pertains to the source region. The inner part of each finger is pinned to low temperature values by the contacts, which results in a fairly complex temperature shape and demonstrates the inherently three-dimensional nature of the electro-thermal behavior.

References

Fig. 2.

Drain Contacts

Source Contacts

Gate

Fig. 3.
Fig. 4.

Fig. 5. $V_{ds} = 5V$
Fig. 6.

Fig. 7.