Electro-Thermal Effects in Mixed-Mode Device Simulation

T. Grasser and S. Selberherr
Institute for Microelectronics, TU Vienna
Gusshausstr. 27-29, A-1040 Vienna, Austria
E-mail: grasser@iue.tuwien.ac.at

ABSTRACT

Due to the ever increasing packaging density of integrated circuits, self-heating and thermal coupling effects become more and more important. For state-of-the-art mixed-mode device simulation the solution of the basic transport equations for the semiconductor devices is directly embedded into the solution procedure for the circuit equations. Compact modeling is thus avoided and much higher accuracy is obtained which is especially true for the temperature dependence of the device terminal characteristics. We review the state of the art in mixed-mode device simulation with particular emphasis placed on self-heating of individual and thermal coupling effects between different devices.

I. INTRODUCTION

Over the last decades numerous powerful circuit simulation programs have been developed. Amongst those are general purpose programs which have been designed to cope with all different kinds of circuits and special purpose programs which provide highly optimized algorithms for, e.g., filter design. General purpose programs can be divided into two categories. Programs belonging to the first category offer a modeling language which can be used to define fairly arbitrary dependences between the circuit elements. The most prominent member of this category is ASTAP [1] which was developed by IBM in the 1970s. To provide the user with a maximum of flexibility, ASTAP generates FORTRAN source files which need to be compiled before execution. The other category consists of programs which only allow for a predefined set of circuit elements and dependences. Although the flexibility is strongly diminished, this approach allows for a much faster execution and a compact, highly optimized simulator kernel. The most prominent member of this category is SPICE which was developed at the University of Berkeley [2].

Circuit simulation programs have in common that the electrical behavior of the devices is modeled by means of a compact model that is an analytical expression describing the device behavior. Once a suitable compact model is found, it can be evaluated in a very efficient way. However, this task is far from being trivial and many complicated models have been developed. Even if the behavior of the device under consideration can be mapped onto one of the existing compact models, the parameters of this compact model need to be extracted. For example, in the case of the BSIM3v3 model [3] for short-channel MOS transistors more than 100 parameters are available for calibration purposes, the identification of which is obviously a cumbersome task. Similar arguments hold for other available MOS transistor models as the EKV model [4], [5] and the Philips MM9 model [6]. If the device design is known and not modified, these parameters need to be extracted only once and can be used for circuit design provided the accuracy of the models is sufficient. When there is need to optimize a device using modified geometries and doping profiles the compact model parameters have to be extracted for each different layout as many of these parameters are mere fit parameters without any physical meaning.

The electrical behavior of the devices can either be measured or simulated. When per-
forming a device optimization, fabricating and measuring each optimization step would be very expensive. Hence, device simulators became more and more popular, e.g., DESSIS [7], GALENE [8], MEDICI [9], MINIMOS [10], and PISCES [11]. These device simulators solve the transport equations for a device with given doping profiles and a given geometry. The transport equations form a highly nonlinear partial differential equation system which cannot be solved analytically. Numerical methods have to be applied to calculate a solution by discretizing the equations on a suitable simulation grid. The data obtained from these simulations can be taken to extract the parameters of the compact model.

Altogether, this subsequent use of different simulators and extraction tools is cumbersome and error-prone. To overcome these problems several solutions have been published where a device simulator was coupled to SPICE [12], [13]. This is again problematic when considering the communication between two completely different simulators. On the other hand some solutions were presented where circuit simulation capabilities were added to a device simulator [14]. However, the restrictions imposed are so severe that circuits containing more than a few distributed devices cannot be properly dealt with.

The examples in this paper were simulated using the device simulator MINIMOS-NT which has been equipped with full circuit simulation capabilities with the only limitation being the amount of available computer resources [15], [16], [17]. MINIMOS-NT is a general purpose device simulator developed as the successor of MINIMOS [18].

Due to the ever increasing packaging density of integrated circuits, self-heating and thermal coupling effects become more and more important. These alter the device performance by inclusion of thermal diffusion currents and by the temperature dependence of the physical parameters, e.g., for the band edge energies, recombination rates, and mobilities. Unfortunately, such electro-thermal problems are very difficult to handle for several reasons. First, the heat spreading volume is normally much larger than the electrically active area and extends to several 100 nm³. Secondly, thermal effects are real three-dimensional effects which cannot easily be approximated by two-dimensional cross-sections as is the case for purely electrical problems. The two-dimensional thermal boundary conditions are difficult to formulate basically for two reasons: the Neumann boundary condition in the third dimension for the heat flux equation causes an overestimation of the temperature which can exceed 100 % and thus renders the results unusable except for first principle estimations [19]. This can also lead to severe numerical problems since the overestimation leads to local lattice temperatures during solver iteration, that exceed the validity of the lattice temperature dependent material models. Secondy, the thermal boundary conditions are determined by the thermal resistors at material transitions as much as by the bulk properties. As these unknown thermal resistivities of industrially relevant materials, such as e.g. glue or thermal bumps, have to be verified by experiments, a self consistent fitting procedure and various assumptions will always be part of application oriented simulations. Thus the thermal boundary conditions have to be determined with respect to aspects normally not included in device simulation such as neighboring devices or chip mounting.

For RF devices, even more aspects have to be considered, since for large signal use, part of the DC power is converted into microwave power, so not all DC power leaves the devices through the thermal contacts as assumed by the DC self-heating model [19].

With mixed-mode capabilities at hand devices can be characterized by their performance in a circuit as a function of transport models, doping profiles, mobility models, device temperatures, etc. This is of fundamental importance when investigating the behavior of modern submicron devices and non-mainstream devices like Heterostructure-Bipolar-Transistors (HBTs) [17] or High-Electron-Mobility-Transistors (HEMTs) [20], [21], [22] where compact models are not so far developed. Furthermore, when the devices
are scaled down, non-local effects become more and more pronounced which can alter the device behavior significantly. These effects cannot be handled by scaling the parameters of compact models.

II. THERMAL SIMULATION

The standard way of treating temperature effects in semiconductor devices and circuits is based on the assumption of a constant device temperature which can be obtained by a priori estimates on the dissipated power or by measurements. However, in general this a priori assumed dissipated power is not in accordance with the resulting dissipated power. Furthermore, devices may be thermally coupled resulting in completely different temperatures than would be expected from individual self-heating effects alone. This is of special importance as many circuit layouts rely on this effect, e.g., current mirrors and differential pairs [23]. Therefore, the temperature must not be considered a constant parameter, but must be introduced as an additional solution variable [24], [25], [26], [27].

Thermal coupling can be modeled by a thermal circuit [23], [28] (cf. Fig. 1). The topological equations describing a thermal circuit are similar in form to Kirchhoff’s equations and the branch relations map to familiar electrical branch relations. The electrical compact models have been extended to provide the device temperature as an external node (e.g., [24], [29]).

To account for self-heating effects, traditionally the lattice heat flow equation [30] is solved.

\[
\begin{align*}
\text{div} \ S_L &= H - \rho_L \cdot c_L \cdot \frac{\partial T_L}{\partial t} \\
S_L &= -\kappa_L \cdot \text{grad} \ T_L
\end{align*}
\]

(1)

(2)

\( S_L \) is the lattice heat flow density and the coefficients \( \rho_L, c_L, \) and \( \kappa_L \) denote the materials mass density, specific heat, and thermal conductivity, respectively. \( H \) is the generated local heat density and can be modeled by an expression given by Adler [31]

\[ H = \text{div} \left( \frac{E_C}{q} \cdot J_n + \frac{E_V}{q} \cdot J_p \right) \]  

(3)

\( E_C, E_V, J_n, \) and \( J_p \) are the conduction band edge energy, the valence band edge energy, the electron and hole current densities, respectively. (3) accounts for both Joule heat generation and recombination heat. However, the influence of thermo-electric effects (Seebeck, etc.) is neglected in both (1) and (3). For more details on this subject see the excellent paper by Wachtuka [32]. Of course, temperature dependent models are used for all physical parameters needed in the device equations, e.g., for the band edge energies, recombination rates, and mobilities [15], [33].

Two simple thermal contact models are commonly used. The first model implements an isothermal contact by simply setting the lattice temperature at the interface points equal to the contact temperature (Dirichlet boundary condition).

\[ T_L = T_C \]  

(4)

The second model is of Cauchy type, it considers a thermal contact resistance at the contact boundary and determines the normal component of the flux. Thus, the expression for the thermal heat flow density \( S_L \) at the contact reads

\[ \mathbf{n} \cdot S_L = \frac{T_L - T_C}{\rho_{th}} \]  

(5)

with \( \rho_{th} \) being the thermal contact resistivity and \( \mathbf{n} \) the normal vector to the surface. The thermal contact conductance \( G_{th} \) is related to the thermal resistivity \( \rho_{th} \) by \( G_{th} = A/\rho_{th} \) with \( A \) as the contact area. For more sophisticated models see e.g. [32].

III. EVALUATION OF THE BOUNDARY CONDITIONS

As an example device we consider a SiGe HBT structure as investigated in [34] with an additional thermal contact added at the right
side of the device assuming a mirror symmetry of the device structure. This contact models the thermal heat flow along the chip surface. The resulting device structure is shown in Fig. 2. Also shown is the region of maximum heat generation which is in the base-collector space-charge region were the maximum collector-emitter voltage drop occurs. It must be kept in mind that the ratio of the heat flows over these four contacts is determined by the design of the chip and the environment it is used in. For the following we assume properly designed heat sinks which drain the generated heat mainly towards the collector and along the chip surface.

For the simulation, unless otherwise noted, the following thermal contact conductances were used: $G_{th}^B = G_{th}^C = 10$ mW/K and $G_{th}^C = G_{th}^{Chip} = 50$ mW/K. Simulated temperature cross sections through the center of the device are shown in Fig. 3 and Fig. 4 for the isothermal and the resistance contact model, respectively, with $V_{CE} = 3.5$ V. Fig. 3 shows the temperature distribution for different base-emitter voltages $V_{BE}$ whereas for Fig. 4 $V_{BE} = 1.0$ V was used and $G_{th}^C = G_{th}^{Chip}$ were varied. For $V_{BE} = 1.0$ V both contact models generate temperature distributions of similar shape but in the case of the resistance

Fig. 2. Geometry and region of maximum heat generation of the example HBT.

Fig. 3. Lattice temperature distribution of an HBT with the isothermal contact model for different bias voltages.

Fig. 4. Lattice temperature distribution of an HBT for different thermal contact conductances.

Fig. 5. Heat generation distribution of an HBT for different thermal contact conductances.
contact model the temperature is shifted by an offset which exponentially depends on $G_{th}^{Chip}$. For $G_{th}^{Chip}$ as small as 10 mW/K no meaningful solution is found as the lattice temperature exceeds 600 K which inhibits a successful simulation since the temperature dependent models leave their range of validity. Furthermore, other effects like impact ionization become important for higher collector-emitter voltages. Although $V_{BE} = 1.0$ V is quite high it must be pointed out that even for lower bias conditions the same situation occurs for improper choice of $G_{th}^{Chip}$.

These investigations show that the simulation result is very sensitive to the contact resistances. Furthermore, it follows that the isothermal model must be used with great care and only when the exact contact temperatures are known.

In Fig. 5 the heat generation inside the device is shown for different values of $G_{th}^{Chip}$. As the current density remains approximately constant within this cross-section, the maximum of the heat generation is located at the base-collector space charge region where the electric field is maximal. As $V_{CE} = 3.5$ V was assumed which is quite moderate, even higher heat generation rates can be expected for power circuits. Although the final values may give reasonable temperature distributions, during iteration the bias voltages of a device in a circuit may vary considerably and can easily exceed $V_{BE} = 1.5$ V and $V_{CE} = 20$ V. This situation can occur during mixed-mode simulation of circuits with large supply voltages and cause excessive problems when simulating fully-coupled electro-thermal systems especially as measured values for $G_{th}^{Chip}$ are in the range 1–10 mW/K.

In Fig. 6 the temperature distribution for different base-emitter voltages is shown generated with a quite large value of 100 mW/K for $G_{th}^{Chip}$. All these figures indicate, that the heat generated inside the device accumulates because it cannot be drained off by the thermal contacts. Thus, the local temperature rise inside the device is much smaller than the temperature rise induced by the contact model. This is especially true for this example as silicon is a good thermal conductor and large temperature gradients are not likely to occur inside the device.

It might therefore not be necessary to perform a fully consistent self-heating (SH) simulation by solving the lattice heat flow equation. Instead, we could use a global self-heating model (GSH) and calculate the dissipated power as

$$P = \sum_C I_C \cdot V_C$$  \hspace{1cm} (6)

with $I_C$ and $V_C$ being the contact currents and voltages. The spatially constant lattice tem-
temperature is modeled as

$$T_L = T_C + P \cdot R_g$$  \hspace{1cm} (7)$$

with $R_g$ being the global thermal resistance. This model is commonly used in compact modeling (e.g., [24], [29]). However, when applying this expression to mixed-mode device simulation, each device is modeled at a distinct device-specific temperature which has a significant impact on device performance. With this approach it is thus possible to make use of all temperature dependent physical parameter models, e.g., mobility, thus significantly increasing simulation accuracy.

The GSH model gives only two additional unknowns ($T_L$ and $P$) compared with the pure electrical system. $R_g$ should be equal to the effective thermal contact resistance plus an equivalent resistance of the device which can be approximated as

$$R_g = R_{th}^\text{eff} + \frac{w}{A \cdot \kappa}$$  \hspace{1cm} (8)$$

with $w$ being the average distance of the thermal contact to the region where the heat is generated and $A$ being the average area of the section connecting the junction with the thermal contact. $\kappa$ is the thermal conductivity of the underlying material which shows a strong temperature dependence [30] and must therefore be evaluated at an average temperature value. Of course, this formula is far too simple to give exact results and it is better to consider $R_g$ a mere fitting parameter only roughly approximated by (8).

As another example we simulated a state-of-the-art Double Base Bipolar Junction Transistor (DBBJT). For this device a proper model for the polysilicon Emitter contact is of fundamental importance to achieve good accuracy. We implemented the model given in [35]. The output characteristic is shown in Fig. 9 where the GSH model with $R_g = 800 \text{ K/W}$ and the SH model delivered the same results (within 2%). Furthermore it can be seen that without consideration of self-heating effects, the simulation deviates significantly from the measurement. For an average operating point the convergence properties of the GSH model are similar to that of the T300 model whereas the SH model takes 20 % more iterations and twice as long in terms of CPU time (see Table I). This advantage of the GSH model becomes even more significant for higher biases. In addition, the GSH model is approximately as robust as the T300 model and we were able to find a solution even for very high device temperatures were the SH model already failed.

IV. EXAMPLE

Thermal effects are of fundamental importance for the chip design of integrated circuits. Typical operational amplifiers (OpAmps) can deliver powers of 50–100 mW to a load, and
as the output stage internally dissipates similar power levels the temperature of the chip rises in proportion to the dissipated output power \cite{28,36}. As the transistors are very densely packed, self-heating of the output stage will affect all other transistors. This is especially true as silicon is a good thermal conductor, so the whole chip tends to rise to the same temperature as the output stage. However, small temperature gradients develop across the chip with the output stage being the heat source. The temperature coefficient of the junction voltage for forward-biased pn-junctions is known to be approximately $-2 \text{ mV/K}$, that is to obtain the same current a smaller junction voltage is needed. These temperature gradients appear across the input components of the OpAmp and induce an additional input voltage difference which is proportional to the output dissipated power.

The complete \mu A709 \cite{23}, \cite{37} as shown in Fig. 10 has been simulated considering thermal interaction between the input and the output stage. It should be noted that all transistors have been simulated by numerical means and not with analytical models. This circuit is of special interest as it is one of the SPICE benchmark circuits given in \cite{2} (without thermal feedback). The DC transfer characteristic has been calculated with and without thermal interaction. Consideration of thermal interaction was first performed by solving the SH model for the transistors $T_1$, $T_2$, $T_9$ and $T_{15}$ and by assuming a thermal network which provides for the thermal coupling of the devices as shown in Fig. 11.

The thermal conductances were assumed to be $G_1 = G_2 = 2 \text{ mW/K}$ and $G_9 = G_{15} = 10 \text{ mW/K}$ while the coupling mismatch was modeled by $G_{1,9} = G_{1,15} = G_k = 10 \text{ mW/K}$ and $G_{2,9} = G_{2,15} = G_k \cdot (1 - \Delta)$ with $\Delta = 0.9$ being the mismatch parameter which is proportional to the temperature gradient across the input transistors \cite{28}. In addition, thermal interaction was considered by using the GSH model in substitution for the lattice heat flow equation.

No problems occurred during the solution of the purely electrical system. Even the consideration of thermal interaction using the GSH model caused no problems. The solution of the fully coupled electro-thermal equation system

\begin{table}
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Method & System-Size & CPU & It \\
\hline
T300  & 4956 & 33 s & 26 \\
GSH   & 4958 & 37 s & 27 \\
SH    & 6738 & 62 s & 31 \\
\hline
\end{tabular}
\caption{Computational details for the example DDBJT ($V_{CE} = 5 \text{ V}, I_B = 8 \text{ \mu A}$).}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{fig10}
\caption{Schematic of the \mu A709 OpAmp.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{fig11}
\caption{Thermal equivalent circuit used to simulate thermal interaction for the \mu A709 OpAmp.}
\end{figure}
using the SH model was possible with a properly designed iteration scheme which works as follows: in the first block the thermal quantities were ignored until an electrical solution was found. In the second block, the lattice temperature was added to the solution vector without considering the coupling effects caused by the node temperatures. This was also found to be advantageous when stepping through the DC transfer curve hence this block was also used for the consecutive steps. After having established a proper temperature distribution inside the devices for the new voltage boundary conditions, the complete equation system can be solved.

The DC transfer characteristic was calculated by stepping \( \varphi_{in} \) from \(-1 \text{ mV}\) to \(1 \text{ mV}\) with \(\Delta \varphi_{in} = 20 \ \mu \text{V}\). From SPICE simulations the open-loop gain of the \(\mu\)A709 was known to be approximately 35000 so for each step of \(\Delta \varphi_{in}\) a step of 0.7 V could be expected for \(\Delta \varphi_{out}\) which is quite large. However, no convergence problems occurred until \(\varphi_{out}\) approached 0 V. This was the most critical part of the simulation and several step reductions for the input voltage were necessary for the SH model. Details of the simulations are summarized in Table II (for a Linux Pentium II 350MHz workstation).

The DC transfer characteristic is shown in Fig. 12 with the obvious humps in the SH models resulting from thermal feedback effects. The GSH model perfectly fits the results obtained by the more complex SH model. In Fig. 13 the open-loop voltage gain \(A_v\) is shown demonstrating the dramatic impact of thermal coupling. The thermal conductances assumed in this simulation were very optimistic and an even stronger impact of thermal coupling has been published [25], [26]. For stronger coupling, even the sign of the open-loop voltage gain may change and cause the OpAmp to become unstable [36].

The maximum temperature and the contact temperature of the output stage are shown in Fig. 14. Information about the maximum temperature is lost for the GSH model, though. As can be seen, self-heating inside the transistor plays only a minor role at these current levels. However, the power dissipated inside the device heats up the NPN transistor due to the resistive thermal boundary condition which obstructs the heat flow out of the transistor. This is the reason for the excellent results obtained by the simple GSH model. The PNP transistor has a \(\beta\) of only approximately
Fig. 14. Maximum and contact temperature of the output transistors $T_9$ and $T_{15}$ during the DC transfer characteristic for both self-heating models.

10 and comparable current levels have been obtained by increasing the emitter area of the transistor ($W_{PNP}/W_{NPN} = 5$). Hence the locally generated heat density $H$ is even smaller than for the NPN transistor and the temperature drop inside the device is negligible thus resulting in nearly no loss of information for the GSH model.

A similar situation occurs for the input transistors $T_1$ and $T_2$. As they are biased with $I_C = 20 \mu A$ only self-heating is negligible and the contact temperature resembles the heat transferred from the output stage thus again resulting in negligible loss of information for the GSH model. As non-symmetric thermal conductivities have been assumed the temperature of $T_1$ is always slightly higher than the temperature of $T_2$. The maximum temperature difference of the input transistors $T_1 - T_2$ was found to be only 22 mK. Even this small temperature difference has such a strong impact on the output characteristic due to the high gain of the circuit.

CONCLUSION

We have investigated the impact of a computation time efficient approach to cover self-heating effects on device and circuit performance. It was shown in a realistic example that self-heating is dominated by the resistive thermal boundary conditions. Thus, the lattice heat flow equation can be substituted by a global self-heating model with nearly no loss of accuracy in the electrical terminal characteristic. This observation is of fundamental importance in the case of mixed-mode device simulations where thermal-coupling effects dramatically increase the complexity of the problem. Using this approximation the problem can be solved in considerably less time with reasonable accurate inclusion of thermal effects. The benefits provided by this approach can be even better exploited in three-dimensional device simulations as there the reduction in the number of unknowns is obviously even more significant.

ACKNOWLEDGMENT

This work is supported by: INTEL Corporation, Santa Clara, USA; LSI Logic Corporation at Milpitas, USA; Philips at Eindhoven, The Netherlands; Infineon Technologies at Munich, Germany; and Sony at Atsugi, Japan.

REFERENCES


