Mixed-Mode Device Simulation

T. Grasser and S. Selberherr

Abstract—In 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. We review the state of the art in mixed-mode device simulation. In addition we present recent achievements, in particular, techniques for convergence acceleration and methods for dealing with electro-thermal problems. Much emphasis is put on the examples section to demonstrate the value and usefulness of the proposed techniques.

I. INTRODUCTION

Over the last decades numerous powerful circuit simulation programs have been developed. Amongst these 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 performing 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 used to calculate a solution by discretizing the equations on a suitable simulation grid. The data obtained from these simulations can be used 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. MINIMOS-NT is a general purpose device simulator developed as the successor of MINIMOS [15].

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, etc. This is of fundamental importance when investigating the behavior of mod-
ern submicron devices and non-mainstream devices like Heterostructure-Bipolar-Transistors (HBTs) \[16\] or High-Electron-Mobility-Transistors (HEMTs) \[17\], \[18\] 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. This cannot be handled by scaling the parameters of compact models.

II. CIRCUIT SIMULATION

Several different methods have been published for the description of the circuit equations. However, nearly all circuit simulators employ methods based either on the nodal approach (NA) \[19\], \[20\], \[21\] or the tableau approach \[22\]. Methods based on the NA enjoy large popularity due to its ease of use. However, the basic NA only allows for current-defined branches. Voltage-defined branches can be introduced without extending the formulation by the use of gyrators \[23\], \[24\]. To properly account for voltage-defined branches the modified nodal approach (MNA) has been proposed which allows for the introduction of arbitrary branch currents \[25\].

III. 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 \textit{a priori} estimates on the dissipated power or by measurements. However, in general this \textit{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 \[26\]. Therefore, the temperature must not be considered a constant parameter, but must be introduced as an additional solution variable \[27\], \[28\], \[29\], \[30\].

Thermal coupling can be modeled by a thermal circuit \[26\], \[31\] (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. For distributed devices MINIMOS-NT solves the lattice heat flow equation \[32\] to account for self-heating effects. This is of course far more accurate than assuming a spatially constant temperature in the device and estimating the dissipated power by Joule-heat terms alone as is done for the compact models. To provide a connection to an external thermal circuit arbitrary thermal contacts are defined.

IV. DEVICE SIMULATION

The vast majority of today's routinely performed device simulations are based on a numerical solution of the basic semiconductor equations which include drift-diffusion current relations \[32\], \[33\], \[34\], \[35\]. The efficiency of this numerical device model allows its extensive use in device optimization.

A device of a modern ULSI circuit is characterized by large electric fields in conjunction with steep gradients of the electric field and of the carrier concentrations. Under these conditions, the accuracy of the widely used drift-diffusion model becomes questionable. More sophisticated device models, such as the hydrodynamic transport model \[36\], \[37\], \[38\], \[39\], \[40\], \[41\], \[42\] overcome these limitations. However, the increased physical rigor of a model comes at the expense of increased computation times. This fact prevented widespread application of the hydrodynamic model in the past, and probably in the near future. This is especially true for mixed-mode simulations which inherently suffer from large simulation times and poor convergence properties. Thus, the necessity of using the hydrodynamic model should be checked by comparison with drift-diffusion simulation results. However, for this comparison to deliver useful results, several prerequisites must be met, the most important of them being that both transport models must deliver similar results under homogeneous situations \[43\], \[44\].

V. MIXED MODE SIMULATION

Several works dealing with circuit simulation using distributed devices have been published so far \[12\], \[13\], \[14\], \[45\]. Most publications deal with the coupling of device simulators to SPICE. This results in a two-level Newton algorithm since the device and circuit equations are handled subsequently. Each solution of the circuit equations gives a new operating point for the distributed devices. The device simulator is then invoked to calculate the resulting currents and the derivatives of these currents with respect to the contact voltages. In \[13\] a method was proposed which was termed full-Newton algorithm. However, this approach is very similar to the two-level method proposed in the same paper thus it will be termed “quasi” full-Newton. The difference to the two-level Newton lies in the fact, that the device simulator only performs the first step of the Newton iteration and returns the result to the circuit simulator. Both approaches are easy to implement as only marginal changes in both simulators are required.
The circuit simulator acts as a server which controls the device simulator. At each Newton iteration of the circuit, an input deck for the device simulator has to be generated and the device simulator has to be called to calculate currents and conductances. The main advantage of this approach is that the device and circuit simulator are decoupled and special device simulators may be used for different problems.

The other approach is called full-Newton algorithm as it combines the device and circuit equations within one single equation system. This equation system is then solved applying Newton's algorithm. In contrast to the two-level Newton and the quasi full-Newton algorithm where the device and circuit unknowns are solved in a decoupled manner, here the complete set of unknowns is solved simultaneously. In MINIMOS-NT an approach similar to [14] is used. The capability to solve circuit equations was added to the simulator kernel. This allowed for assembling the circuit and the device equations into one system matrix which results in a real full-Newton method. There is no need to explicitly calculate the derivatives of the contact currents with respect to the contact voltages as the contact currents are solution variables which simply gives ±1 as a derivative in the constitutive relations.

However, the benefits gained from using the numerous existing SPICE compact models must not be neglected. As SPICE has a well defined and documented interface, it is, in principle, straightforward to implement a similar interface in the combined circuit-device simulator.

A comparison of these different architectures is shown in Fig. 2. In Fig. 2a the device simulator acts as a client to the circuit simulator whereas in Fig. 2b the device simulator is extended with circuit simulator capabilities and can reuse circuit simulator models on demand.

VI. Convergence

The system of equations which has to be solved for mixed-mode device simulation is non-linear and extremely sensitive to small changes in the solution variables. While the semiconductor equations are difficult to solve themselves the situation becomes even worse when using dynamic mixed-mode boundary conditions. To solve these equations the Newton method is used which is known to have quadratic convergence properties for an initial-guess sufficiently close to the final solution. However, such an initial-guess is hard to construct for both the distributed quantities inside the device and the circuit equations. Hence methods have to be found to enlarge the region of convergence to succeed even with a poor initial-guess. This is achieved by applying suitable damping schemes. One of the most popular damping schemes has been published by Bank and Rose [46], [47]. In MINIMOS-NT a purely heuristic method is used which takes the exponential relation between the potential and the carrier concentration into consideration [48]. This method provides similar convergence properties to the method of Bank and Rose without costly evaluation of damping parameters.

Especially important is a reliable method to obtain a DC operating point which is needed as a starting point for a subsequent transient analysis or a static transfer characteristic. Transient simulations are far better conditioned as the time derivatives provide main-diagonal entries and act as a natural damping. As the solution of the last timestep provides a good initial-guess it is normally possible to obtain convergence for a sufficiently small timestep. Although the conditioning of the equation system does not change for DC transfer analysis the last solution again provides a good initial-guess. In case the system fails to converge for a given step the step can normally be reduced in such a way to obtain convergence. Hence the following discussion will focus solely on DC operating point calculation.

To the best knowledge of the authors no useful damping scheme for mixed-mode has been published so far. Only in [12] it was stated that the change of the node voltages was limited to a user-specified value which is in the range of 2·V_T. This is, as pointed out in the very same paper, far from being optimal as it requires a large number of iterations for larger supply voltages. E.g., for the OpAmp circuit simulated in the examples section the supply voltages are ±15 V, hence it takes at least 15/0.05 = 300 iterations to build up the supply voltages without even considering the effect of non-linearities. Furthermore it is stated in [12] that a solution can only be obtained for an initial-guess as close to the solution as ±0.2 V for forward-biased junctions.

These restrictions of mixed-mode simulations seem to be generally accepted nowadays. Experiments with a new method delivered promising results for small circuits, the main field of application of mixed-mode simulations. This method is based on the idea, that the distributed devices should be carefully embedded into the rest of the circuit during evolution of the operating point. Similar observations were made by Ho et al. [49] for FET circuits using compact models. They proposed to shunt a resistor of 3 kΩ at the source and
drain during the first three Newton iterations to stabilize the coupled system and to slightly decouple the device from the circuit equations. This approach has been extended by introducing an iteration dependent conductance $G_k$ between each device node and ground. The following purely empirical expression for $G_k$ delivered very satisfying results

\begin{align}
G_0 &= 10^{-2} \, S \\
G_{\text{min}} &= 10^{-12} \, S \\
G_k &= \max\left( G_{\text{min}}, G_0 \cdot 10^{-k/\kappa} \right) \\
\kappa &= 1.0 \ldots 4.0
\end{align}

with $k$ being the iteration counter. It is worthwhile to note that the algorithm worked equally well with $G_{\text{min}} = 0$ for the simulated circuits. However, this expression is purely empirical but unfortunately any attempt to use a more rigorous expression based on norms of the quantities did not work satisfactorily.

Using this new technique, solutions could be found for several typical analog and digital circuits starting from the zero initial-guess for the node voltages and charge neutrality assumptions for the semiconductor devices within 20-50 iterations which is a comparable effort to SPICE which uses compact models.

VII. EXAMPLES

A. Five-Stage CMOS Ring Oscillator

A five-stage CMOS ring oscillator circuit is shown in Fig. 3. For both the NMOS and the PMOS transistors a device width of $W = 1 \, \mu m$ was assumed. Normally, to achieve equal noise margins, a ratio of $W_p/W_n \approx 2.5$ is used to compensate for the poorer performance of the PMOS transistor \cite{50}. To model the influence of the interconnect circuitry, an additional load capacity of 5 $\text{fF}$ was used. To force the circuit into a predefined initial state, the input voltage $\varphi_{\text{in}}$ of the first inverter was set to zero during operating point calculation.

Two different ring oscillators have been simulated, one with long-channel transistors ($L_G = 2 \, \mu m$), the other one with short-channel transistors ($L_G = 0.2 \, \mu m$). For the long-channel transistors, the simulation results obtained with the drift-diffusion and hydrodynamic transport models agree so closely, that in the graph no differences are visible (cf. Fig. 4). The simulation results for the short-channel devices are shown in Fig. 5. Here, the differences between the transport models are significant. This is due to the larger currents resulting from the hydrodynamic transport model as the charging and discharging times of an inverter chain are inversely proportional to the drain currents. The simulated inverter delay times are $\tau_{\text{DD}} \approx 30 \, \text{ns}$ and $\tau_{\text{HD}} \approx 26 \, \text{ns}$ giving a difference of about 15 %. For single devices the hydrodynamic currents are approximately 30\% and 5\% higher for the NMOS and the PMOS transistor, respectively. The average of these
values (17.5%) closely corresponds to the simulated delay time difference of 15%.

B. Five-Stage CML Ring Oscillator

A current mode logic (CML) gate is an emitter coupled logic (ECL) gate stripped of the emitter-follower [50], [51]. The gain of a single stage without load can be approximated by assuming a simple Ebers-Moll model for the transistors [52] to be approximately $-5$. When considering an inverter chain consisting of 5 CML inverters as shown in Fig. 6, the total gain occurring at the last output node is $(−5)^5 = 3125$. With such a high gain, the circuit is too sensitive to the voltage changes occurring during iteration such that no solution can be found without a proper initial-guess using conventional techniques. However, using the shunt conductance technique with $\kappa = 4$ a DC operating point was easily obtained with only 34 iterations. As for the CMOS ring oscillator there is no unique operating point for the closed-loop and one of the node voltages had to be fixed to force the circuit into an initial state from which oscillations can start. Oscillations start immediately with a frequency $f_{DD} = 6.8$ GHz for the drift-diffusion and $f_{HD} = 10.6$ GHz for the hydrodynamic model which gives a relative difference of 36% for the drift-diffusion model (Fig. 7). This is due to the velocity overshoot which occurs in the base-collector space charge region which cannot be modeled using a drift-diffusion transport model. The current levels are approximately equal in both cases.

C. Electro-Thermal Analysis of a Complete OpAmp

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 [31], [53]. 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 $-2$ 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 [26], [54] as shown in Fig. 8 has been simulated considering thermal interaction between the input and the output stage. This circuit is of special interest as it is one of the SPICE benchmark circuits given in [2] (without thermal feedback). The DC transfer characteristic has been calculated with and without thermal interaction. Consideration of thermal interaction was done by solving the lattice heat flow equation for the transistors $T_1, T_2, T_5$ and $T_{15}$ and by assuming a thermal network which provides for the thermal coupling of the devices as shown in Fig. 9. The thermal conductances were assumed to be $G_1 = G_2 = 2$ mW/K and $G_6 = G_{15} = 10$ mW/K while the coupling mismatch was modeled by $G_{1,8} = G_{1,15} = G_8 = 10$ 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 [31].

The solution of the fully coupled equation system is possible with a proper iteration scheme. A small change in the output voltage during iteration causes a large change in the collector current of the conducting output transistor. The dissipated power changes which influences the temperature distribution inside the output transistor. This modified power alters the base-
emitter voltages of the input transistors which produces a change in the base-emitter voltages of the output transistors. As all these coupling mechanisms are highly non-linear a special iteration scheme is used. 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. However, as the condition of the transient problem is much better, this block is not used for transient simulation. Only after a proper temperature distribution inside the devices has been established for the new voltage boundary conditions, can the complete equation system be used.

However, as the simulation failed very frequently for too large steps of the input voltage an additional failure criterion was added. When the step of the input voltage was too large it caused oscillations in the solution which, due to the strong non-linearities, blew up the lattice temperatures. This took approximately 30 iterations which were very expensive in computational terms as each iteration took approximately 20 - 200 seconds depending on the condition of the system matrix. So this event had to be detected as soon as possible. It was found that an abnormal behavior of the potential update norm \( \| \Delta \|_\infty \) was a good indication of starting oscillations. Hence, whenever \( \| \Delta \|_\infty \) was larger than approximately \( 10^2 \cdot V_T \) after 10 iterations or whenever \( \| \Delta \|_\infty \) exceeded \( 10^3 \cdot V_T \), the iteration was canceled. Furthermore, the number of iterations was limited to 30.

The DC transfer characteristic was calculated by stepping \( \varphi_{in} \) from \(-1 \, \text{mV} \) to \(1 \, \text{mV} \) with \( \Delta \varphi_{in} = 20 \, \text{mV} \). 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 were necessary for both the pure electrical and the thermal simulation. The size of the system matrix was 37177 and 40449 for constant temperature and thermal simulation, respectively, and the simulation took 9 and 25 hours on a Linux Pentium II 350MHz workstation. For the thermal simulation the conditioning of the system matrix was found to be very poor and several step reductions were necessary.

The DC transfer characteristic is shown in Fig. 10 with the obvious humps resulting from thermal feedback effects. In Fig. 11 the open-loop voltage gain \( A_v \) is shown and 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 [28], [29]. For stronger coupling, even the sign of the open-loop voltage gain may change and cause the OpAmp to become unstable [53].

The maximum temperature and the contact temperature of the output stage are shown in Fig. 12. It is obvious that the 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 in accordance to the commonly used assumption that the transistor can be modeled by a power source alone. The PNP transistor has only a \( \beta \) of approximately 10 and comparable current levels have been ob-
tained by increasing the emitter area of the transistor ($W_{NP} / W_{PN} = 5$). Hence the locally generated heat density $H$ is even smaller than for the NPN transistor and the temperature drop inside the device is negligible.

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. As unsymmetric thermal conductivities have been assumed the temperature of $T_1$ is always slightly higher than the temperature of $T_2$. The maximum temperature difference $T_2 - T_1$ was found to be only $-22$ mK. Even this small temperature difference has a strong impact on the output characteristic due to the high gain of the circuit.

**CONCLUSION**

The state of the art in mixed-mode device simulation has been reviewed considering different approaches. A method has been proposed which allows for operating point calculations without an initial-guess for medium-sized circuits. Several important examples were presented to demonstrate the value and usefulness of mixed-mode device simulation.

**ACKNOWLEDGMENT**

This work is supported by: Hitachi Ltd. at Tokyo, Japan; INTEL Corporation, Portland, USA; LSI Logic Corporation at Milpitas, USA; Philips at Eindhoven, The Netherlands; Siemens Corporation at Munich, Germany; and Sony at Atsugi, Japan.

**REFERENCES**


