Transistor Level Modeling for Analog/RF IC Design

Edited by
W. Grabinski, B. Nauwelaers
and D. Schreurs

Springer

TRANSISTOR LEVEL MODELING FOR ANALOG/RF IC DESIGN

Transistor Level Modeling for Analog/RF IC Design

Edited by

WLADYSLAW GRABINSKI
Geneva Modeling Center, Freescale, Switzerland

BART NAUWELAERS
K.U. Leuven, Belgium

and

DOMINIQUE SCHREURS
K.U. Leuven, Belgium

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-4555-7 (HB)
ISBN-13 978-1-4020-4555-4 (HB)
ISBN-10 1-4020-4556-5 (e-book)
ISBN-13 978-1-4020-4556-1 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

Printed in the Netherlands.

TABLE OF CONTENTS

FOREWORD

Among many great inventions made in the 20th century, electronic circuits, which later evolved into integrated circuits, are probably the biggest, when considering their contribution to human society. Entering the 21st century, the importance of integrated circuits has increased even more. In fact, without the help of integrated circuits, recent high-technology society with the internet, cellular phone, car navigation, digital camera, and robot would never have been realized. Nowadays, integrated circuits are indispensable for almost every activity of our society.

One of the critical issues for the fabrication of integrated circuits has been the precise design of the high-speed or high-frequency operation of circuits with huge number of components. It is quite natural to predict the circuit operation by computer calculation, and there have been three waves for this, at 15-year intervals. The first wave came at the beginning of the 1970s when LSIs (Large Scale Integrated circuits) with more than 1000 components had just been introduced into the market. A mainframe computer was used for the simulation, and each semiconductor company used its own proprietary simulators and device models. However, the capability of the computer and accuracy of the model were far from satisfactory, and there are many cases of the necessity of circuit re-design after evaluation of the first chip.

The second wave hit us in the middle of 1980s, when the EWS (Engineering Work Station) was introduced for use by designers. At that time, most of the simulation tools were already provided by software vendors and standard device models for public use were being established. The simulation of circuits became considerably more accurate and the amount of re-design was significantly reduced. The third wave started to flood us at the beginning of this century, when the PC provided sufficiently high performance for circuit simulation. We are facing the front of the third wave now.

The situation for device models for circuits has changed very much during the past 20 years. Ten years ago, the model was kept in strict secrecy within each

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, vii–viii.

semiconductor company, in most cases. Now the semiconductor companies adopt open public standard models or even common model parameters in order to provide a familiar design environment to their customers. Recently, very accurate and complex models have been required to cope with the need to design extremely high-speed logic circuits with ultra-small transistors – sometimes even with an SOI substrate. The characteristics of ultra-small CMOS devices are quite different from those of older larger transistors and the models become very complicated. Also, the macroscopic treatment of large number of logic devices with hardware description languages such as VHDL-AMS or Verilog-A becomes very important, with tremendous increases in integration.

Another aspect is that the market for RF integrated circuits has become very large, and there are strong demands for an accurate RF model for CMOS and HBTs. Traditionally, modeling of RF devices was very difficult, because of the accuracy required not only for the first derivative of the I-V characteristics, but also for the third derivative. In addition, an accurate three-dimensional model of the substrate is essential for precise RF simulation of active and also passive components. The substrate model is also important for noise simulation, which is a key element in RF devices. Corresponding to the accuracy and complexity of the model required, the expression of the model has wide variety; empirical expression, analytical expression, table look-up expression, and numerical expression obtained by numerical simulation.

The importance of compact modeling for circuits is becoming bigger and bigger in the third wave, and we expect to see great progress. This book includes some of the recent important advances in compact modeling. It is our hope that this book will be useful for designers and modeling scientists facing the front of the third wave.

Hiroshi Iwai
Tokyo Institute of Technology

December 1, 2005

INTRODUCTION

Wladek Grabinski, Bart Nauwelaers and Dominique Schreurs

The accuracy of the integrated circuit analysis performed in contemporary design flows is directly correlated to the quality of its fundamental components – the models. To ensure on-time delivery of these models, characterization and model generation must be rapid and precise. To be able to take full advantage of the new semiconductor technologies, the designers have to update their CAD tools regularly with precise definitions of the new device models that can be implemented into circuit simulators and design flows. The models must preferably be physics-based to account for complex dependences of the device properties on dimensions and other process variables. The model parameters are derived from measurements and characterization of the devices. For RF CMOS (bulk and SOI) and compound technologies, both modeling and characterization are challenging tasks that will be especially emphasized in this book.

This book is aimed at radio frequency (RF)/analog and mixed-signal integrated circuit (IC) designers, computer-aided design (CAD) engineers, semiconductor physics students, as well as wafer fab process engineers working on device, compact model level. We can summarize the goals of the book as follow:

- to give the reader a consistent introduction to the main steps of compact model developments, including advanced 2/3D process and device simulations, consistent and accurate MOSFET modeling founded on the physical concepts of the surface potential, charge-based modeling, empirical modeling of small and large signal device behavior, and modeling approaches that are based on linear as well as non-linear measurements;
- to illustrate the impact of device-level modeling on IC design using selected examples;

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, ix–xiii.

– to provide a detailed insight into modeling and design flow automation based on high-level behavioral languages, i.e. VHDL-AMS and Verilog-A.

We have structured this book to cover the key aspects of compact model developments, showing consistent flow of the implementation and dissimilation as well as its standardization tasks. Following that organization, the book is divided into ten main chapters:

In the first chapter of this book, D. Donoval *et al.* introduce 2/3D process and device simulation as an effective tool for better understanding of the internal behavior of semiconductor structures. Process simulations are used to create a virtual device with geometry and properties identical to the real structure, and such basic technology steps as ion implantation, diffusion, epitaxial growth, oxidation, deposition and etching are presented. Then numerical 2/3D device simulations are performed. The complete simulation flow is illustrated by three advanced examples: a bipolar transistor, a CMOS inverter structure and a power vertical DMOS transistor multi-cell structure. These kinds of virtual device structures created by 2/3D process and device simulations are often used as initial inputs for compact model development and validation.

Bulk CMOS models make up the main stream of the compact models. Next, two chapters discuss two concepts of the physics-based models for CMOS devices.

R. van Langevelde *et al.* present PSP: an advanced surface-potential-based MOSFET. The PSP compact model jointly developed by Philips Research and Pennsylvania State University is based on fundamental physics (the surface potential approach) over the entire MOSFET device operating regime. Such effects as gate leakage, noise, non-quasistatic (NQS) and quantum-mechanical effects, which become increasingly important with the downscaling of CMOS technology, are physically modeled within PSP and have been verified experimentally. The model also provides a better description of high-frequency behavior. The PSP model enables improved simulations of a wide class of circuits including analog/RF modules that are important in the mobile communication technology and other advanced designs. The PSP compact model is supported by professional software environments, including Verilog-A, which allow it to be directly coupled to many popular circuit simulators. The PSP model has been submitted to the Compact Model Council (CMC) as a candidate for standardization.

M. Bucher *et al.* present EKV3.0: an advanced charge-based MOSFET transistor model which is design-oriented towards next-generation CMOS technologies and IC designs. Historically, the development of the EKV model is driven by the needs of analog IC designers. This chapter presents the physical foundation of the EKV charge model, which is itself based on a surface-potential analysis. The basic charge modeling approach allows not only

physically consistent and accurate modeling of current, terminal charges and noise, but also offers a unique set of suitable expressions for hand-calculation of analog/RF circuits. The fully-featured EKV3.0 compact MOST model for circuit simulation is presented and validated using advanced RF application examples down to sub-100 nm CMOS technologies. Finally, the parameter extraction procedure and implementation in the Verilog-A language are briefly discussed.

The next two chapters describe empirical models, and also include information on measurement techniques used for model extraction or creation.

Reliable measurements are a prerequisite for any sensible device modeling work, in particular for RF applications where the silicon or III–V compound material-based device models are required to predict their subtlest behavior. D. Schreurs focuses on MOSFET modeling using direct high-frequency measurements. After explaining the theoretical background of two high-frequency modeling approaches, the author discusses the main characterization steps, i.e. linear and non-linear vector measurements and the importance of de-embedding, as well as equivalent circuit and behavioral modeling. Both linear and non-linear measurement-based modeling approaches are explained and their different implementations are illustrated by examples.

I. Angelov discusses empirical FET models. Experimental static current, S-parameter and capacitance characteristics are linked with small and consistent large signal equivalent circuit modeling, leading to an empirical FET model. This creates a basis for reviewing Standard and Extended Curtice Models, the Materka-Kacprzak Model, the Triquint Model, the EESOF Model, and the Chalmers FET Model. The author also shows an extended empirical model to incorporate physical phenomena such as thermal effects and dispersion.

The following three chapters bring some specific physical aspects into the modelling arena: SOI with its special substrate build-up, effects of very small dimension MOSFETs, and quantum effects that are observable in some circuits.

Silicon-on-insulator (SOI) CMOS technologies offer exceptional advantages not only for digital designs but also for RF, low-GHz telecommunication and microwave IC designs. B. Parvais and A. Siligaris present an empirical approach to modeling the SOI MOSFET nonlinearities. The analytical model is introduced to describe the nonlinear behavior of the SOI device from DC to RF coherently, and to account for the dispersive character of some physical phenomena, such as floating body (FB) effects in the SOI device. The simulations of a new model were validated by measurements and explained by a simple analytical model, based on the Volterra series approach.

Some of the models might not properly describe some physical effects presented in aggressively down-scaled CMOS technologies. As the MOSFET models are critical for reliable RF designs, new physical effects must be incorporated and alternative modeling techniques must be proposed. N. Itoh focuses on and describes some insufficiently modeled phenomena in the recent small

geometry MOSFETs, i.e. mobility degradation due to STI stress and channel noise enhancement due to hot carrier effects. His model accounts for physical effects associated with STI stress, scalable parasitic components and channel thermal noise allowing the reduction of both the cost and design period of advanced RF/analog IC design.

F. Felgenhauer *et al.* discuss incorporation of parasitic quantum effects in classical circuit simulations. The performance of a state-of-art CMOS device is influenced by an increasing number of parasitic effects associated with recent down-scaling of integrated semiconductor devices. Beside semi-classical parasitic effects and leakage currents such as sub-threshold current, DIBL and GIBL, further parasitic effects of quantum mechanical origin must be included in device modeling. The discussion covers the physics and the simulation of coherent charge transport with a successful attempt to include quantum effects in high-level circuit simulations such as SPICE. The simulation model developments are illustrated by three different circuit examples, which explicitly exhibit the influence of quantum effects on circuit functionality.

The two final chapters provide detailed insight into how modeling and design flow automation can be supported and enhanced by analog hardware description languages (AHDLs) such as VHDL-AMS and Verilog-A.

C. Lallement *et al.* present the capabilities of the VHDL-AMS hardware description language for compact model development. The chapter is a case study and shows that VHDL-AMS can be successful used for implementation of such models as EKV 2.6 and MM11 MOSFETs. The authors also show that the basic models can be easily enhanced to include major physical effects like self-heating, extrinsic aspects and quantum effects, since the VHDL-AMS language naturally supports multi-domain. VHDL-AMS is not limited to single device compact modelling but also can be used to describe innovative integrated devices, like Micro-Opto-Electro-Mechanical Systems (MOEMS) integrating different application-field parts on the very same chip (e.g. mechanical, electrical, thermal, and fluidic parts). Similarly to Verilog-A, discussed in the next chapter, application of VHDL-AMS to compact modeling is an attempt to standardize the compact modeling development environment.

B. Troyanovsky *et al.* present Verilog-A, a behavioral language for compact modeling of MOSFET developments, as a platform-independent software tool. The authors introduce Verilog-A, a general-purpose modeling language, by examples guiding the reader through language elements, operators, functions and structure, with particular emphasis on the constructs important to the compact model developer. The recent Verilog-A release of the language standard has added several features of interest to compact model developers. The main language extensions are discussed in the chapter. It is important to note that several academic and industrial model development groups, i.e. PSP and EKV teams, now use Verilog-A as a main and a platform-independent language of their development methodology.

From this summary of the contents of the book, the reader can see that a broad overview of modelling techniques in the MOS arena is described by a select group of authors. Very fine contributions regarding the best compact models are complemented with equally good work regarding measurement-based modeling. Additionally a number of specific topics on SOI, small devices, quantum effects and hardware description languages (VHDL-AMS, Verilog-A) further increase the usefulness of this book.

Bringing together such a notable group of authors is a very visible result of the ongoing effort of the MOS-AK group, behind which one of the editors, W. Grabinski, is a driving force, to bring all European researchers in the advanced MOS field together and to keep them talking about their mutual research interests, and more specifically about the modeling aspects of MOS devices and circuits.

It was just after the MOS-AK workshop organized in September 2004 at the University of Leuven that the co-operation between the publisher and the editors to create this book was initiated. The editors would like to thank the authors of the various chapters and the publishers' staff for bringing this project to a successful conclusion.

Chapter 1

2/3-D PROCESS AND DEVICE SIMULATION

An effective tool for better understanding of internal behavior of semiconductor structures

Daniel Donoval[1], Andrej Vrbicky[1], Ales Chvala[1], and Peter Beno[2]
[1]*Department of Microelectronics FEI, Slovak University of Technology in Bratislava, Ilkovicova 3, 812 19 Bratislava, Slovakia*
E-mail: daniel.donoval@stuba.sk
[2]*ON Semiconductor Slovakia*

Abstract: 2/3-D numerical process and device simulation is presented as an extremely useful tool for the analysis and characterization of fabrication processes and corresponding electro-thermal behavior of semiconductor structures and devices standing alone and/or coupled in integrated circuits. In the introductory part of this chapter, a brief description is given of the basic features, processes, and structures implemented in the numerical process and device simulation. Visualization of the internal properties (electrical, thermal, optical, magnetic, and mechanical) allows comprehensive analysis of the critical regions and weak points of the analyzed structures. The presented examples illustrate the potential, power and beauty of numerical simulation of processes and devices for the identification and analysis of the behavior of parasitic devices that exist as inevitable parts of active devices and which degrade the normal operation and reliability of integrated circuits. Commercially available TCAD process and device simulators with verified calibrated complex electro-physical models, advanced numerical solvers securing stable calculations, and user friendly interactive environment provide a unique insight into the internal operation of the analyzed structure. They can be efficiently used for comprehensive physical interpretation of experimentally obtained results and/or particularly for prediction of the properties and behavior of new semiconductor structures and devices as well as for further development and optimization of new technologies and fabrication steps.

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, 1–27.

Key words: process and device simulation; structure (mesh) definition; boundary conditions; electro-physical models; steady state and transient simulation; bipolar and CMOS technology; DMOS technology – power devices; parasitic devices; latch-up effect; electro-thermal interaction.

1. Introduction

Enormous advances in the microelectronics technology with an exponential growth of the complexity and speed following the Moore law [1] and SIA Roadmap [2] are required to secure a continuous development of new technologies, structures, devices, circuits and systems. The better understanding of the electro-physical behavior and potential of new structures and devices with dimensions scaled down to deep submicron range and operating at their physical limits put stringent requirements on modeling and simulation. Since trial manufacturing of highly dense IC with minimal dimensions of individual devices in deep submicron region costs a great deal, modeling and simulation play an increasingly important role in the development and prediction of the properties of modern technologies. By means of simulation, microscopic physical phenomena and effects occurring on very small length scales and in very short time periods can be visualized in macroscopic dimensions and, thus, perceivable to our eyes and mind.

Over the past thirty years, Technology CAD (TCAD) has evolved into a well-accepted branch of the global electronic design automation environment (EDA) characterized for example by a recent acquisition of TCAD tools developer and vendor ISE AG Zurich by Synopsys. Single simulators for process simulation, device simulation, parameters extraction and circuit simulation are integrated by interactive user friendly graphical environments and provide the virtual wafer fab GENESISe-ISE [3] and VWF of Silvaco [4] allowing cost and yield estimation as well as comprehensive parametric analysis of semiconductor processing. Introduction and integration of new physical models for thermo-opto-electro-mechanical effects into advanced simulators enables the simulation of the properties and behavior of microtransducers and very complex micro-opto-electro-mechanical systems (MOEMS). Comprehensive surveys of different physical models, methods of mathematical treatment, features of data compatibility and handling, their visualization and examples of applications of numerical process and device simulation can be found in a large number of books and proceedings [5–7].

The increasing on-chip circuit and system integration allowed by continuing miniaturization of individual semiconductor devices, which are approaching their physical limits, generates a strong pressure on a better understanding of the electro-physical behavior of individual semiconductor structures integrated in IC technology [8]. Design of advanced semiconductor devices with minimum

dimensions at nm scale working in high frequency applications, however, calls for new advanced complex physical models including quantum-mechanical effects for a wide variety of semiconductors, insulators and metals (Si, SiGe, GaAs and other III–V compounds, high k-oxides, silicides) [9]. Mixed mode device and small signal circuit simulators including numerical simulation of 2/3-dimensional structures predicting their behavior, properties and reliability are unavoidable tools of any research team working in the development and optimization of new fabrication processes. There is a continuous need for new experts with complex knowledge and skills who will be able to solve the global problems [10, 11].

In spite of that, most system engineers working in IC design laboratories with EDA tools work on higher abstraction levels with limited knowledge of the internal behavior of individual devices including their parasitic components. Therefore the main aim of this chapter is a presentation of the potential, power and beauty of numerical process and device simulation with its unique insight into the internal semiconductor structure operation for a better understanding of the integrated circuit behavior under various stress conditions in different environments. The reader who is interested in the state of the art numerical process and device simulations including the most advanced physical models with quantum-mechanical effects for deep submicron structures and devices is referred for example to [12] for more details.

A brief description of process and device simulators, their structures, required input parameters, used physical models, format and visualization of output data, and potential applications will be presented. The given examples will characterize the big potential of numerical process and device simulation for a unique insight into the analyzed structure and for identification and analysis of the behavior of parasitic devices that are inevitable parts of almost all active devices in various technologies of IC's.

2. Process Simulation

The behavior and properties of all semiconductor devices are defined by their three geometrical dimensions and concentration profile of impurities. The main goal of process simulation is to model a virtual device with geometry and properties identical with the real structure. The lateral dimensions which specify the active parts of the devices are defined by lithography masks, while the vertical depth and concentration of active impurities depend on the used fabrication processes. Each fabrication process can be modeled usually by a set of partial differential equations (PDE's), which can be solved either analytically and/or numerically. The advanced physical models with calibrated parameters characterizing individual fabrication steps are integrated into the process simulators. As technology development continues, the need for new more precise

process models increases. Continuous calibration of their parameters is based on the best correlation of simulated results with experimental data acquired on special test structures by analytical tools such as secondary ion mass spectrometry (SIMS).

Numerical solutions exploit iterative numerical solvers which calculate the structure properties in a defined region with properly defined boundary conditions. Dense grids with a high number of nodes, where the individual unknowns and properties are defined, provide a higher accuracy, the tradeoff being a longer elapsed time and memory. Therefore, adaptive grid generation in curved regions with steep profiles of physical entities is a necessity particularly for more dimensional simulations. The output results are mostly represented by 2D doping profiles with a 1D cross section, which provides information about the concentration of impurities in selected cross sections in horizontal or lateral dimensions. While some years ago 2D models and solutions were fully sufficient, nowadays only 3D simulation can take into account the global complexity and variety of various phenomena occurring in miniaturized deep sub μm and nano-structures. However, due to the enormous requirements on the computing resources and computing time (full 3D simulation of complete technological process is in general still beyond the capabilities of most today's software tools and computers) they will not supersede the 2D simulations in the near future. To solve the tradeoff between the grid with an increased number of nodes and computation time and memory requirements the simulators allow simulating one half of a symmetrical structure, which is then reflected across the selected boundary.

The current commercially available simulators provide an interactive environment with high a degree of flexibility for input commands, implement advanced physical models with calibrated parameters and numerical solvers with efficient meshing for robust and stable simulation.

The input commands of individual steps make accessible all parameters which characterize the real fabrication processes. They comprise:

Ion implantation – the process by which impurity atoms are implanted into active parts of the substrate material defined by a mask with a given dose, energy and tilt angle, which prevents creation of impurity tails due to the channeling effect. The resulted doping profiles correspond to analytical distribution functions (Gaussian, Pearson, dual Pearson) with tabulated parameters such as the projected range and lateral straggle depending on the collision mechanisms of specific implanted species with the substrate material [13]. If the tables are not available, Monte Carlo simulators for ab initio calculation of interactions of implanted atoms with the substrate atoms can be used [14]. As the projected range in general increases with smaller atoms, BF_2 molecules are used for implantation of shallow junctions to prevent deep penetration of light materials like boron (B). The process of ion implantation creates a big amount of defects and amorphization of Si single crystal occurs when using high doses.

To activate the implanted impurities to the lattice positions and recrystallize the damaged and/or amorphous regions, high temperature annealing should follow the ion implantation process.

Diffusion – is a high temperature process of diffusion of impurities due to the existing concentration gradient, which depends on temperature and time of diffusion, boundary conditions characterizing the surface (interface) concentration of diffusion species at the Si substrate and gas interface. The time and position dependent concentration of impurities are the solution of PDE's (Fick diffusion equations). Various physical models with different levels of complexity depending on the type of impurity (its temperature and concentration dependent diffusion coefficient), point defects and electric field effects implemented in advanced simulators are very well described in [15]. For example, the simplest constant diffusion model which neglects the interactions between the dopants and point defects and electric field effects is used mainly for dopant diffusion in oxides. The pair diffusion model assumes that the gradient of dopant concentration and dopant-defect pairs with the electric field are the driving force of diffusion in active Si regions predefined by the mask. As processing proceeds through various annealing cycles and the concentration gradient exists, the dopants diffuse and redistribute through the structure, therefore the temperature budget should be minimized to ensure very steep and shallow doping profiles for miniaturized structures and devices.

Epitaxial growth – is a growth of single crystalline Si layers on top of the Si substrate at temperatures slightly lower than the melting point. The thickness of the growing epitaxial layer is characterized by the growth rate and time. Various impurities, different in concentration or species from substrate impurities, can be incorporated into the epitaxial layer. As it is a high temperature process, redistribution of impurities occurs at the interface due to the concentration gradient.

Oxidation – is a process of growth of thermal silicon dioxide (SiO_2) at the silicon surface depending on temperature, time and oxidation ambient characterizing the diffusion of oxidants from the gas-oxide interface to Si-SiO_2 interface and its reaction with Si. As the process of thermal oxidation is accompanied by volume expansion, which invokes strong mechanical stresses and materials motion, the ramping up and down temperature cycles with slow temperature changes are used to prevent structure damage. Due to various segregation coefficients of impurities, segregation of dopants occurs at the interface.

Deposition and Etching – are the processes of deposition and etching of different layers (insulators, metals, poly Si). The deposition may be isotropic, anisotropic, polygonal and fill step. The etching means removing of material which is in contact with gas and may be also isotropic, anisotropic and directional. The thickness of a deposited and/or etched layer is defined by the mask and growth/etching rate and time. As the simulated region (volume) is changed, remeshing of the analyzed structure is required.

The input file for the 2D simulation of 0.18 μm NMOSFET with a lightly doped drain in DIOS [15] contains the following commands and parameters:

```
(1)   TITLE("180nm_NMOS")
(2)   # Initial definitions
(3)   grid( x=(-0.4, 0.4) y=(-10.0, 0.0), nx=2)
(4)   substrate (orientation=100, element=B, concentration=5.0E14, ysubs=0.0)
(5)   replace (control(maxtrl=9, refineboundary=-6, refinejunction=-7)
(6)   #Start simulation of Process Steps
(7)   implant (element=B, dose=5.0E13, energy=300keV, tilt=0)
(8)   diff (time=8, temper=900, atmo=O2 )
(9)   deposit (material=po, thickness=180nm) ;poly gate deposition
(10)  mask (material=re, thickness=800nm, x(-0.09, 0.09)) ;poly gate pattern
(11)  etching (material=po, stop=oxgas, rate(aniso=100)) ;poly gate etch
(12)  etching (material=ox, stop=sigas, rate(aniso=10))
(13)  etching ()
(14)  implant (element=As, dose=4.0E14, energy=10keV, tilt=0) ;LDD implantation
(15)  deposit (material=ni, thickness=60nm) ;nitride spacer
(16)  etching (material=ni, remove=60nm, rate(a1=100), over=40)
(17)  etching (material=ox, stop=(pogas), rate(aniso=100))
(18)  implant (element=As, dose=5E15, energy=40keV, tilt=0) ;N+ implantation
(19)  diff (time=@rta_time@sec, temper=1050, atmo=N2) ;final RTA
(20)  mask (material=al, thick=0.03, x(-0.5, -0.2, 0.2, 0.5)) ;metal contacts
(21)  save (file='180nm_nmos', type=DFISE) ;save final structure
```

The results of numerical process simulation by DIOS-ISE are presented in Figure 1. The generated grid with adapted denser grid points in a curved and steep profile region related to 2D doping profile is shown in Figure 1a.

Corresponding 1D doping profile in A-A cross section designated in a is shown in Figure 1b. The influence of different thermal budget on the lateral distribution of N-type impurities and corresponding shortening of channel length can be clearly seen.

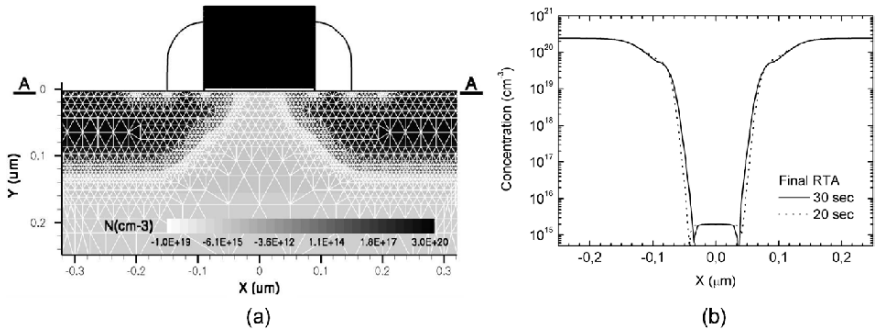


Figure 1. Simulated (a) 2D doping profile with mesh definition, (b) 1D doping profile in A-A cross section for different process temperature budgets.

The simulated results – distribution of dopants in Si are stored in formatted data files and visualization tools are used for quick presentation of the obtained 1D and particularly 2D concentration profiles.

Another important resulted parameter of process simulation is mechanical stress which may induce defects or damage at different layers and interfaces and subsequently influence the electrical properties (interface states density, mobility) of the analyzed structure.

3. Device Simulation

The main goal of device simulation is to provide electrical steady state, transient and small AC signal behavior and characteristics of the studied semiconductor structures for predictive analyses of the properties of new technologies and devices and simultaneously a unique insight into the internal process and structure operation, thus enlarging the users knowledge and expertise. A real semiconductor device, such as transistors, is represented by a virtual device defined by 2/3D structure (output of process simulator) whose electrophysical properties are discretized onto a nonuniform mesh of nodes. The input files for device simulations contain the types of materials, doping profiles of impurities in the given region associated with the discrete nodes, starting temperature, and properly defined boundary conditions with applied external electrical, optical, mechanical, magnetic, and thermal field. An extensive set of advanced electrophysical models with calibrated parameters which characterize the behavior and various effects present in semiconductor structures and interfaces at various applied stresses are incorporated into the advanced device simulators.

The output electrical characteristics are calculated by numerical solution of a set of partial differential equations.

$$\nabla\varepsilon\nabla\psi = -q\left(p - n + N_d^+ - N_a^-\right)$$
$$\nabla\vec{J}_n = qR + q\frac{dn}{dt} \qquad -\nabla\vec{J}_p = qR + q\frac{dp}{dt}$$

For isothermal simulation, the simplest drift-diffusion model comprises three basic semiconductor equations, which are the Poisson and current continuity equations for electrons and holes with potential ψ, free electron and hole concentrations n and p as unknowns. The mobility of free electrons and holes $\mu_{n,p}$, electric field $-\nabla\psi$, generation-recombination rate R and others are considered as variable parameters. They are dependent on the actual values of individual unknowns and therefore an iterative and coupled mode of solution should be used. The total current J in any point of the analyzed structure is then calculated as a sum of electron and hole currents $J_{n,p}$

$$J = J_n + J_p \qquad J_n = -qn\mu_n\nabla\phi_n \qquad J_p = -qp\mu_p\nabla\phi_p$$

where $\mu_{n,p}$ are the mobilities and $\phi_{n,p}$ are the quasi-Fermi potentials of electrons and holes, respectively.

For analysis of devices in which the self-heating effects are not negligible the non-isothermal simulation using a thermodynamic model [16] should be involved. The thermodynamic model assumes that the electrons and holes (their temperatures) are in thermal equilibrium with the lattice temperature and an additional partial differential equation characterizing the influence of self-heating effects and non-isothermal temperature distribution on structure behavior should be coupled and calculated with three basic semiconductor equations.

With continuous miniaturization of semiconductor devices operating in the deep submicron regime the more complex hydrodynamic model [17] should be used for simulation of state of the art devices. In hydrodynamic or energy balance model six PDE's (three basic semiconductor equations and three energy balance equations) should be solved in the coupled mode. The individual free electron and hole temperatures T_n and T_p not equal to the lattice temperature T_l are assumed and calculated from the energy balance equations.

For improvement of the simulation results, particularly for deep submicron devices the Schrödinger equation, which implements the most physically sophisticated quantization model characterizing the tunneling and other quantum-mechanical effects in analyzed structures, should be calculated self-consistently for a more precise evaluation of the potential and free carriers distribution.

A comprehensive review of advanced electrophysical models which complexly characterize the properties and behavior of semiconductor structures and devices can be found in the user manual of simulator DESSIS [18]. Its user friendly interactive graphical environment allows continuous improvement and modification of models and their parameters.

To enlarge the capability, the most advanced simulators provide a mixed mode support for simulation of single or multiple mesh based structures in a circuit with devices defined by SPICE models. For the transient mode of simulation, the device properties are re-solved at any increment of time.

They in general support different device geometries and contain sophisticated nonlinear solvers for numerical simulation. The mesh of nodes should be optimized for any given device structure and type of simulation to get a desired accuracy and efficiency of simulation. The adaptive mesh generators provide densest meshes in the regions with the high gradients of impurities, potential, high current density and curved structures. For example, the simulation of MOSFET requires a very dense mesh in the channel under the gate oxide interface, particularly in the drain region, where the electric field has its highest value (Figure 2).

The influence of the used model (drift-diffusion, thermodynamic, and hydrodynamic) on the output and transfer characteristics of the $1\,\mu$m and

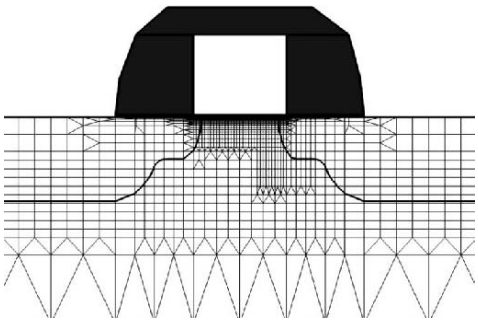


Figure 2. Mesh with non-homogeneous density of nodes of a 0.18 μm NMOSFET.

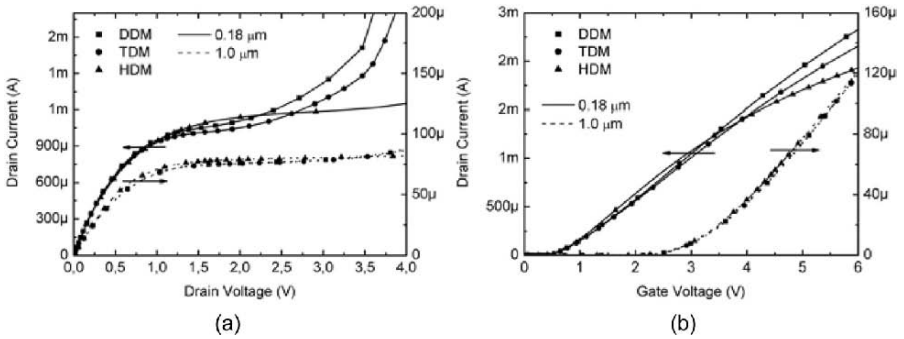


Figure 3. (a) Transfer and (b) output characteristics of a 1 μm and 0.18 μm NMOSFET calculated by drift-diffusion, thermodynamic and hydrodynamic models.

0.18 μm transistors are shown in Figure 3. While the simulated results are similar for all models for 1 μm structure, we can see a big discrepancy for 0.18 μm structure, particularly for a high electrical field, where impact ionization for the drift-diffusion model is overestimated. Therefore the use of the hydrodynamic model for a deep submicron structure is a must.

A unique advantage of process and device simulation is the possibility of simultaneous presentation of output electrical characteristics with visualized internal properties of the analyzed semiconductor structure. Although they can be shown in 1D, 2D or 3D representation, the 2D graphs are most widely used profiles for visualization of different entities. Their correlation with the output characteristics allows analyzing the critical points and regions in the structure depending on the device layout and fabrication design and extract the parasitic devices, which are inevitable parts of many semiconductor structures and devices. Such identified parasitics can be then attributed to the non-standard malfunction behavior of semiconductor devices and IC's. Therefore, reverse engineering based on the interpretation of experimentally obtained data

supported by process and device modeling and simulation is very important not only for the design and optimization of the layout and technology for new devices but also for a better understanding of their properties and behavior.

The 3D simulations require an enormous computer capacity and also 3D visualization of the obtained data, particularly in black & white representation, is not a trivial problem. Therefore, a high degree of user expertise is a must. Nowadays the 3D process and device simulations are still subjects of interest and evaluation in advanced research laboratories, more than the widely applied tools in industrial settings.

An example of 3D thermal simulation for analysis of the temperature distribution in a silicon die is illustrated in Figure 4. Thermal Shut Down (TSD) is a common device in SMART power IC's protecting the whole device against overheating. If the temperature of TSD overcomes a critical value, the power transistor is switched off and no heat is generated any more. The knowledge of the temperature distribution within the die allows the designers to locate TSD close to the hot spots and adjust the appropriate switch off temperature. 3D simulation is necessary to model properly the thermal behavior of a real Si block and 2D and 1D cross sections provide the actual temperature in a selected position.

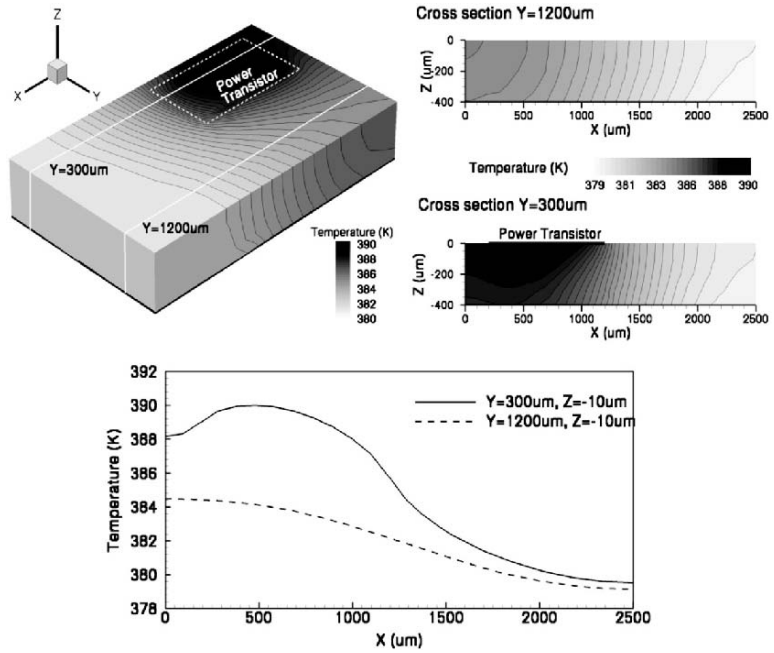


Figure 4. 3D simulation of thermal distribution within a Si block.

4. Examples

Three examples of an efficient use of 2D numerical process and device simulation in the analysis of the output electrical characteristics and extraction of parasitic devices supported by the knowledge of internal properties and behavior of the analyzed structure will be presented.

The first example shows the analysis of a bipolar transistor cell with a buried collector and reverse biased PN junction isolation, where a parasitic lateral bipolar transistor induces a steep increase of the substrate current which contributes to the base current and correspondingly degrades the transistor current gain β.

Analysis of the origin of the latch-up effect and modifications of the fabrication process and design layout of a CMOS inverter structure to increase its robustness against degradation is presented in the second example.

In the third example the complex electro-thermal behavior of a power vertical DMOS transistor multi-cell structure is analyzed, where a parasitic NPN bipolar transistor created under some circumstances generates excessive heat and due to a positive feedback degrades the power transistor.

4.1. Parasitic Lateral Bipolar Transistor in Bipolar Technology

Although the classical bipolar technology is not a mainstream of advanced semiconductor technology, it is still very popular among the designers. The use of 2D numerical process and device simulation for the analysis and interpretation of the measured static *I–V* characteristics of the bipolar NPN transistor and its behavior in the common emitter configuration, namely base, collector, and substrate currents I_b, I_c, and I_s (Gummel plot) and the extracted value of the common emitter current gain β will be presented.

Process simulation by DIOS [15] generating the structure and its doping profile (see Figure 5) and subsequent numerical solution of basic semiconductor equations using the complex physical models implemented in the device simulator DESSIS [18] is used for simulation of static *I–V* characteristics of the bipolar NPN transistor in the common emitter configuration at room temperature (Figure 6). The substrate potential kept at $V_s = -2\,\mathrm{V}$ during all simulations ensures reverse biasing of the N-type collector and P-type substrate isolation junction.

An almost ideal exponential growth is clearly seen of the base and collector currents within many orders of magnitude with corresponding negligible substrate current flowing through a reverse biased PN junction to the substrate. At high values of the base voltage, a sudden super-exponential increase of the substrate current contributes to the total base current and a kink effect in the

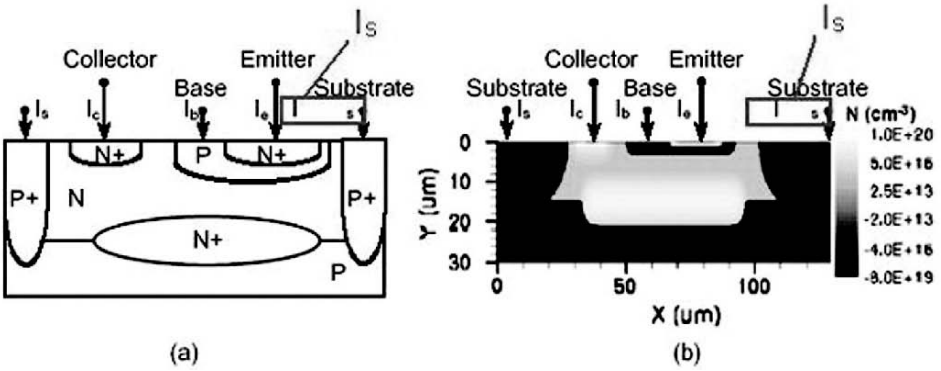


Figure 5. (a) Structure and corresponding (b) 2D doping profile of a bipolar transistor structure cell.

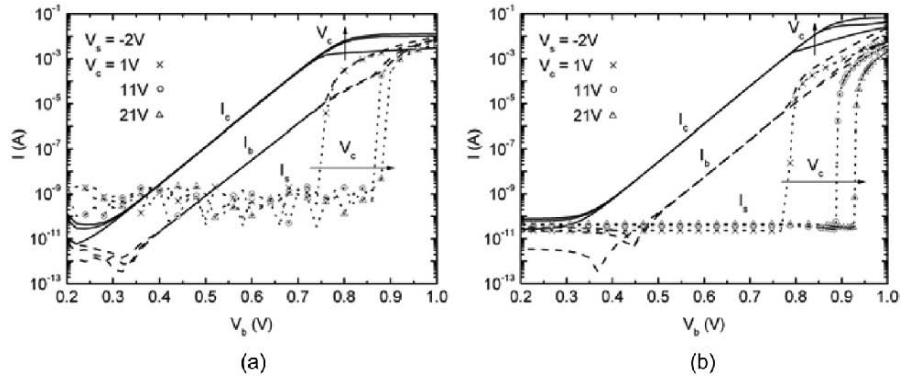


Figure 6. (a) Measured and (b) simulated base, collector and substrate currents I_b, I_c, and I_s in common emitter configuration for different collector voltages $V_c = 1, 11$ and 21 V.

base current is observed. For a proper physical interpretation of this effect, a thorough understanding of the internal behavior of the bipolar transistor cell structure is necessary.

The increasing voltage drop on the series collector resistance decreases the reverse bias of the collector-base junction located on the right side of the analyzed structure far from the collector contact (Figure 7).

For the base voltage of $V_b = 0.86\,\mathrm{V}$ the collector junction is reverse biased in the whole cross section of the analyzed structure. With increasing the base voltage to $V_b = 0.88\,\mathrm{V}$ the collector current and corresponding voltage drop on the series collector resistance increase. There is only a small reverse bias on the collector-base junction, which completely vanishes with a further increase of the base voltage ($V_b = 0.9\,\mathrm{V}$). The collector-base junction which is reverse biased during normal operation of the NPN bipolar transistor becomes open and the holes are injected from the P-type base to N-type collector at the left side of

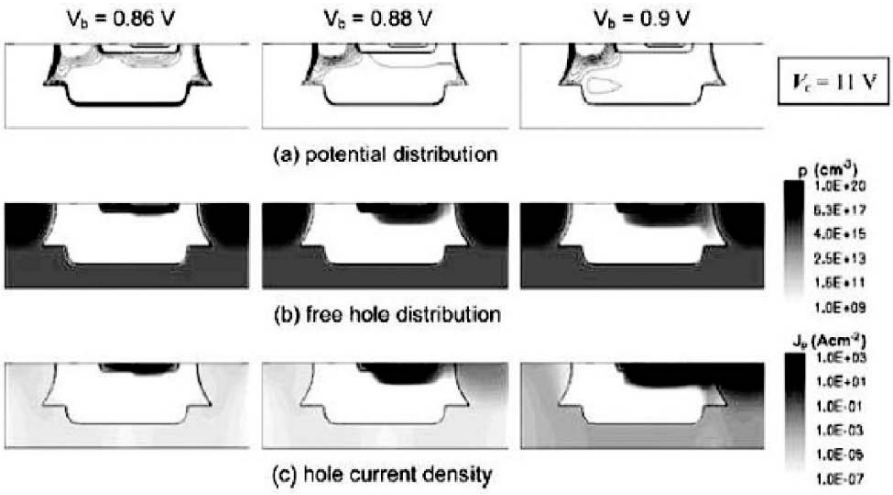


Figure 7. Visualization of the internal properties of a bipolar transistor cell.

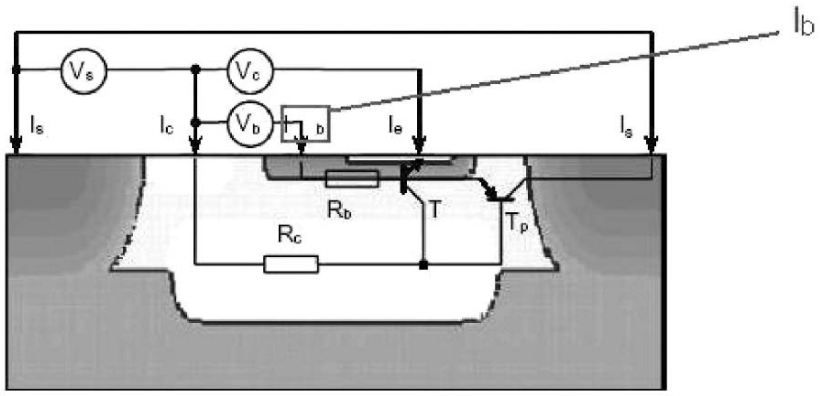


Figure 8. Structure of the bipolar transistor cell and equivalent circuit model for SPICE simulation.

the structure far from the external ohmic contact to the collector (Figure 7b). The holes injected from P-type base to N-type collector are swept by the electric field of the reverse biased junction of the P-type isolation guard ring and the N-type collector and a large hole current starts to flow into the substrate. The described behavior corresponds to the negligible substrate current for $V_b = 0.86\,\mathrm{V}$, its small increase for $V_b = 0.88\,\mathrm{V}$ and finally large increase of the substrate current for $V_b = 0.9\,\mathrm{V}$ (Figure 7c).

Based on the above analysis, the equivalent circuit model for SPICE simulation attributed to the corresponding structure regions was derived (Figure 8) [19]. The bipolar technology with a buried collector and reverse

```
vce 1 0 DC {vc}
Rc 1 2 1000
vb1 4 0 DC 0
Rb 4 3 100
q1 2 3 0 q172
qp 5 2 3 q173
vs 5 0 DC -2
.model q172 npn bf={200+vc*2}
 + is=1e-16 ise=1e-15
.model q173 pnp bf=20
 + is=1e-16 isc=1e-10
.param vc=1
.step param vc list 1 11 21 31
.DC vb1 0 1 0.01
.print dc I (Rb) I(Rc) IC(qp)
.PROBE
```

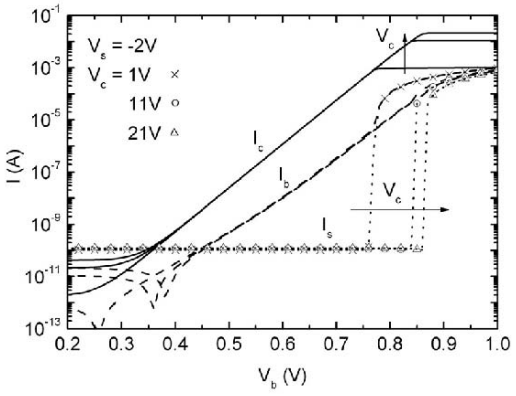


Figure 9. Input netlist and *I–V* characteristics simulated by SPICE.

biased P-type junction isolation may be characterized by a vertical active NPN bipolar transistor with its base and collector series resistances R_b and R_c, and a lateral parasitic PNP bipolar transistor merged with the active transistor. The P-type base and N-type collector of the active vertical transistor create a P-type emitter and N-type base of the parasitic lateral transistor, respectively. The amplifying effect of this parasitic lateral PNP bipolar transistor can be then considered as the origin of the sudden super-exponential growth of the substrate current at a high base voltage, when the large collector current and corresponding voltage drop on the collector series resistance for a given configuration opens the normally reverse biased collector junction of the active bipolar transistor.

The individual components and parameters of the equivalent circuit model (input netlist) for circuit simulation were estimated from 2D device simulation (Figure 9). The obtained *I–V* characteristics simulated by SPICE are in very good agreement with the results of numerical process and device simulation of the corresponding structure of the bipolar transistor as well as with the experimental results, which confirms the validity of the derived model and approach.

4.2. Latch-up Effect in CMOS Technology

The traditional scaling factor $(1/\sqrt{2})$ between successive technology generations allows unprecedented down-shrink of unipolar transistors. which has followed the Moore law [1] for more than 30 years. The key MOSFET design goal is to maximize the transistor speed, and the tradeoff is a relatively high leakage current, corresponding high power consumption and heat dissipation. Also, with MOSFET scaling it will become increasingly difficult to simultaneously

achieve a low sheet resistance for a shallow junction to ensure acceptable series resistances.

Down shrinking of the critical dimensions allows a closer location of NMOS and PMOS transistors. This invokes another problem from which the CMOS technology suffers. Particularly, the big output CMOS inverters and structures for switching applications with an inductive load are sensitive to the so-called latch up effect. We illustrate the origin of latch up on the CMOS inverter structure shown in Figure 10. The two parasitic NPN and PNP bipolar transistors created by N^+-source, P^--substrate and N^--well, and P^+-source, N^--well and P^--substrate, respectively, are clearly seen.

If the output is on logic one and the voltage drop on the series resistance R_n is high enough, the emitter of the parasitic PNP bipolar transistor becomes forward biased and injects holes to the N^--well. These holes are then swept by the electric field of the reverse biased collector junction towards the grounded substrate contact V_{ss} (Figure 11a). The hole current through the series resistance R_p can cause a voltage drop sufficient to open the emitter junction of the parasitic NPN bipolar transistor which injects the electrons to the P^--substrate (base). The injected electrons are then attracted by the electric field towards the N^--well and finally to V_{dd} contact pad (Figure 11b). The electron current increases

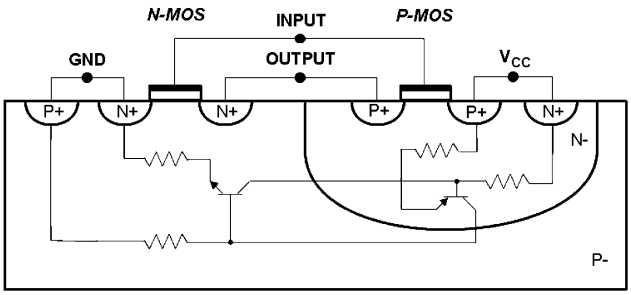


Figure 10. Cross section of CMOS inverter structure A with parasitic bipolar transistors which create a parasitic thyristor.

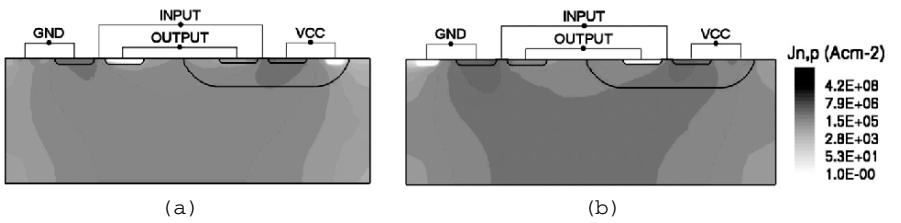


Figure 11. (a) Hole J_p and (b) electron J_n current density in a CMOS inverter structure sensitive to latch up during the trigger current pulse test.

the voltage drop on R_n resistance, which subsequently increases the forward bias of the emitter junction of the parasitic PNP bipolar transistor injecting more holes towards the ground pad V_{ss}. The created positive feedback then leads to a further increase of the total current. A high current continues to flow through structure A also when the trigger pulse is off, which may destroy the device thermally (Figure 12).

In Figure 13 the time dependent response of the output voltage, NMOS and PMOS source currents as well as N^--well and NMOS drain current to input trigger test current impulse $I = 20\,\text{mA}$ are shown. We can clearly see that the output voltage falls down to the thyristor hold voltage and will not recover to the output high value after the trigger impulse is over.

Based on the previous analysis it is clear that the layout design and doping profile should be tuned carefully to protect the device against the latch up.

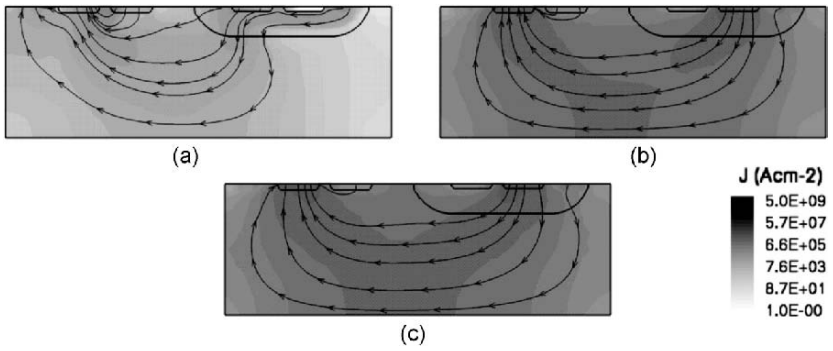


Figure 12. Total current J in a CMOS inverter structure (a) at the beginning (0,1 ms), (b) during (3 ms) and (c) after (7 ms) the trigger current pulse test.

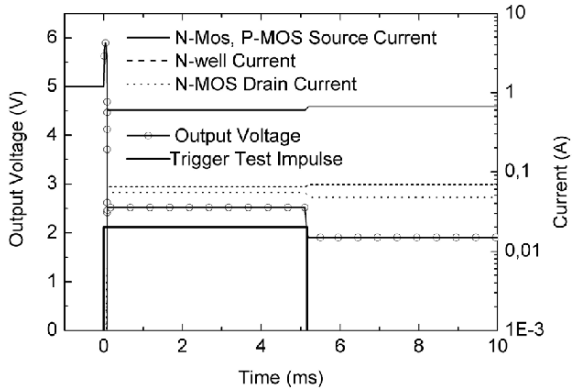


Figure 13. Resulted characteristics of latch up test with trigger current pulse $I = 20\,\text{mA}$ and corresponding output voltage for original structure.

Although there exist different approaches how to avoid or at least minimize the latch up sensitivity [20, 21], two modified structures were analyzed. The analysis followed the test procedure defined by EIA/JEDEC Standard [22], where the devices under test should survive the triggering applied current pulse $I = 100\,\text{mA}$. Interpretation of the obtained results is supported by the 2D numerical process and device simulation with visualized internal properties.

In the first modified structure B we changed the layout and added a P^+-guard ring surrounding the N-channel MOSFET and N^+-guard ring surrounding the P-channel MOSFET (Figure 14a). These guard rings act as additional base contacts of parasitic bipolar transistors and sink the collector currents without a further increase of the open emitter voltage. Although the resistivity of such a structure to the latch up effect is highly improved, it suffers from large area consumption that decreases the density of integration.

To prevent the larger area consumption the concentration profile of impurities was changed in the second modified structure C with the same layout as the original structure A. The latch up robustness was improved by introducing a highly conductive P^{++}-buried layer created on the Si substrate before epitaxial growth of the active layer (Figure 14b).

The resulting characteristics of the latch up test with a trigger current pulse $I = 100\,\text{mA}$ for modified structure C are shown in Figure 15. The output voltage is at its constant high value during the whole test except for two spikes corresponding to the times when the trigger pulse was switch on and off. Similar results were obtained for structure B. It is clear that the resistivity of both structures to latch up was increased considerably and both structures pass the EIA/JEDEC Standard current latch up test.

The internal properties of both structures during and after the trigger pulse are presented in Figure 16. The additional base contacts in structure B sink the hole and electron currents and inhibit creation of the parasitic thyristor. A similar situation is in structure C, where the hole current flows through the highly conductive buried layer and the resulted voltage drop is not sufficient to open and forward bias the NP emitter junction, which prevents formation of

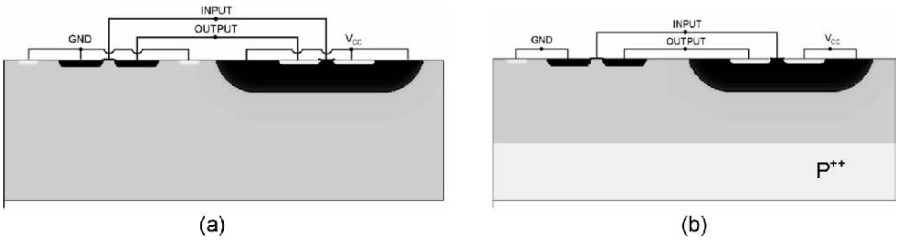


Figure 14. Cross section of the modified structure with (a) guard rings (structure B) and (b) highly conductive buried layer (structure C).

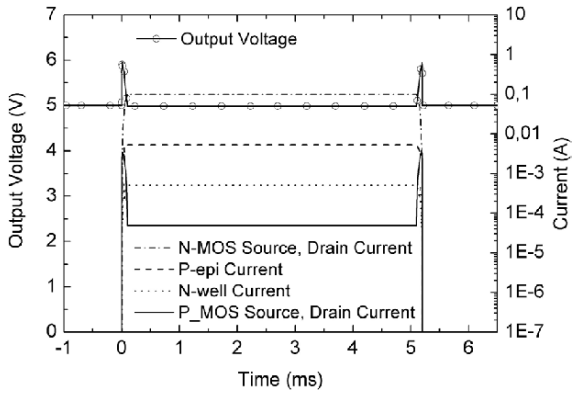


Figure 15. Resulted characteristics of latch up test with trigger current pulse $I = 100\,\text{mA}$ and corresponding output voltage for modified structure C.

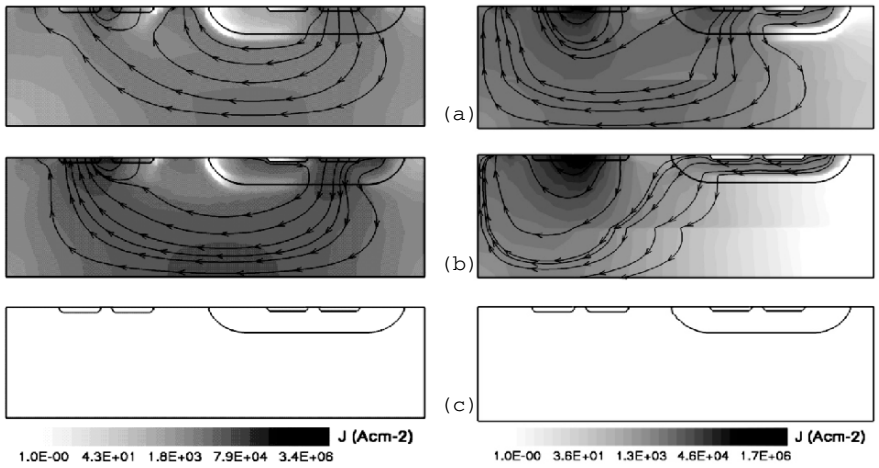


Figure 16. Total current J in a CMOS inverter for structure B (left) and structure C (right) at different time of applied trigger current pulse: (a) $t = 0, 1\,\text{ms}$, (b) $t = 3\,\text{ms}$ and (c) $t = 7\,\text{ms}$.

the positive feedback leading to device failure. We can see that after the trigger pulse the total current drops down to its steady state value for both modified structures.

The presented results of the electrical behavior of three analyzed CMOS inverter structures under latch up test confirm that the 2D process and device modeling and simulation are very efficient, time and cost effective tools for predictive parametric analysis of the sensitivity and robustness of new structures and fabrication processes to the latch up effect.

4.3. Parasitic Bipolar Transistor in Power DMOSFET Technology and its Influence on its Reliability

Many power MOSFETs applications, such as power supplies, DC-DC converters, motor drives and others require devices with a specified breakdown voltage, low on-resistance and high switching speed. For most of these applications, there is a strong demand for devices which should withstand the crucial conditions related to their implementation in switching circuits with an inductive load [23, 24]. Under such extremely harsh switching conditions, the MOSFETs must sustain a great deal of stress without causing destructive failure. The unclamped inductive switching (UIS) condition represents the circuit switching operation for evaluating the "ruggedness", which characterizes the device capability to handle high avalanche currents during the applied stress [25, 26]. We present an experimental analysis of the ruggedness of power DMOSFETs devices. The analysis is supported by the advanced 2D mixed mode device and circuit simulation, which provides a unique insight into the multicell DMOS structure operation and allows to identify the mechanism of current flow through the transistor in its off-state. Finally, creation of a parasitic bipolar transistor and electrothermal behavior of the studied structures are discussed.

The power DMOS transistor contains a large number of individual cells connected in parallel. For our analysis we used numerical simulation of the multicell structure with five adjacent cells (Figure 17). To study the device performance and energy capability, when the transistor is in off state and most of the heat is generated, we set the drain and gate voltages $V_{ds} = V_g = 0$ and assume room temperature $T = 300\,\mathrm{K}$ at the beginning of transient simulation. Hence, for studying the parasitic behavior dependent on self-heating effects, non-isothermal equations using the thermodynamic model must be

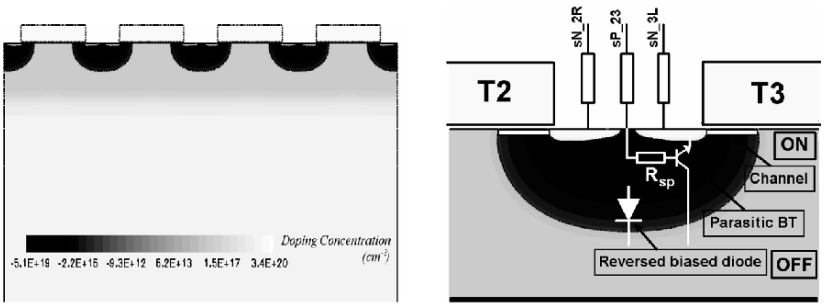


Figure 17. Multicell DMOSFET structure: (a) 2D cross-section, (b) individual cell with highlighted parasitic devices.

incorporated into the device simulation. For transient simulation the drain current I_d was ramped up to 4 mA within 1 μs and the whole simulation period is 100 μs. To obtain realistic electro-thermal characteristics we used a 18 μm wide and 300 μm thick Si block with reflecting boundary conditions at sidewalls and a thermal contact at the device bottom. We modeled the bad cell by a higher series resistance to the P-type well in 2D mixed mode simulation [27]. Such a series resistance characterizes the ohmic contact resistance to the P-well and series resistance of the current path in the P-well as in the real structure the ohmic contact is located in a distance of few μm in the 3rd direction from the analyzed 2D device cross section.

The results of 2D numerical electro-thermal simulation using the thermodynamic model are shown in Figure 18. At the very early stage of the transient simulation ($t = 0.25\,\mu s$) the drain current was homogeneously distributed within all the cells, a slightly smaller current flowed through cell No. 1 due to its higher series resistance $R_{p1} = 2\,k\Omega$ in comparison with other cells, where the resistances were set to $R_{p2-4} = 1.25\,k\Omega$ (Figure 17a). The highest current flowed through the fifth cell with $R_{p5} = 0.625\,k\Omega$. The current flows predominantly through the reverse biased PN junction at the bottom of the P-wells in the avalanche regime (Figure 19a). The highest voltage drop created at the R_{p1} (see inset of Figure 18b) at $t = 0.5\,\mu s$ was sufficient to forward bias the N-emitter and P-well junction which acts as the emitter of a parasitic bipolar NPN transistor. Thus, the conductance of the bad cell was enhanced due to the change of the mechanism and location of the current flow. The original current caused by the avalanche current of the reverse biased PN junction at the bottom of P-well was overtaken by the current of the open parasitic NPN transistor under the channel. Such a cell sinks most of the total current which generated significant Joule heat and resulted in a local temperature growth (Figure 19b). As the avalanche breakdown has a positive temperature coefficient, the drain voltage in the bad

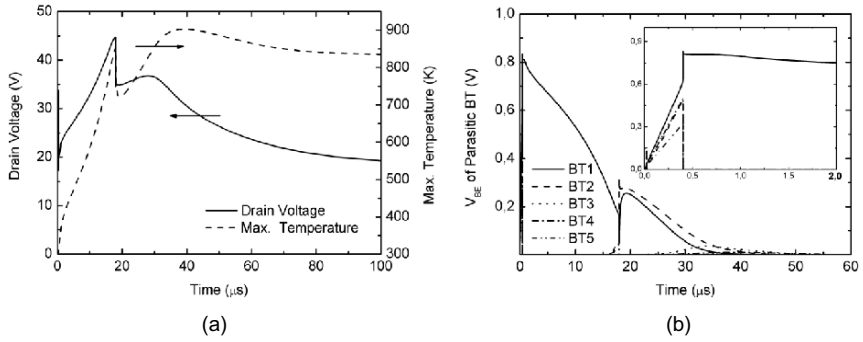


Figure 18. Transient simulation of (a) drain voltage and maximum temperature; (b) inner voltage at emitter junction of parasitic bipolar transistor.

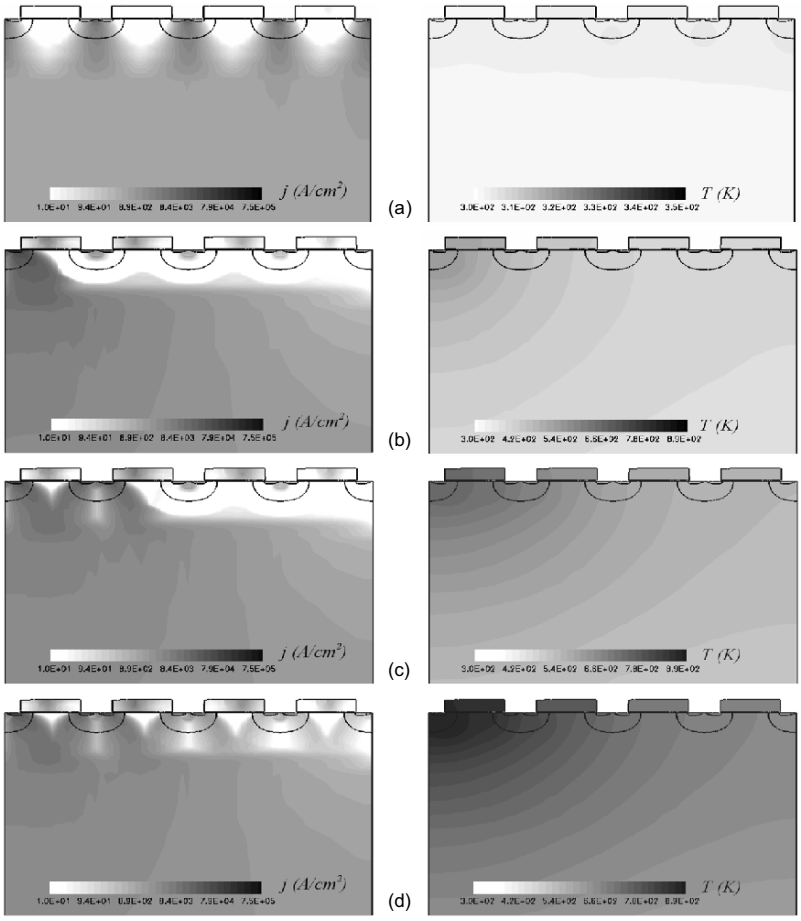


Figure 19. Current and temperature distribution in the analyzed multicell structure at (a) 0.25 μs; (b) 8 μs; (c) 21 μs, and (d) 40 μs.

cell increased further, which was followed by a further temperature growth. The temperature gradient resulted in the heat flow and the closest neighbor cell was heated up above the critical temperature when the second parasitic bipolar transistor in cell No. 2 was opened due to the decrease of the built-in voltage V_{bi} of the emitter-base PN junction with increasing temperature. The non-negligible current started to flow through cell No. 2, which reduced the current via cell No. 1. As a consequence, the generated Joule heat in cell No. 1 decreased and a kink in the drain voltage and maximum temperature can be seen in Figure 2 at elapsed time $t = 21\,\mu$s (Figure 19c). Later, a process similar to that described above took place in cell No. 2, which resulted in a further increase of the drain voltage and local maximum temperature. Due to the heat

transfer the next cells were also heated up and started to conduct higher currents (Figure 19d), which again slightly decreased the total maximum temperature and particularly the drain voltage (Figure 18a). Although oscillations in the drain voltage and maximum temperature appeared during transient simulation of the multicell structure, their physical significance is questionable because the maximum temperature already reached $T \approx 900\,\mathrm{K}$, which should be assumed as a critical temperature for the local destruction of the device [28].

Formation of the parasitic NPN bipolar transistor is a serious concern of the device performance during UIS test. In case the current flowing through an inductance is quickly turned off, the magnetic field induces a counter electromagnetic force (EMF) that can build up surprisingly high potentials across the switch (device under test). The total buildup voltage of this induced potential may far exceed the nominal breakdown voltage $V_{(BR)DSS}$ and energy capability of the transistor, thus resulting in a catastrophic failure [29, 30]. Figure 20 shows a simplified UIS test circuit and corresponding current and voltage waveforms of the tested device under UIS conditions.

The device under test was a conventional vertical DMOS transistor with breakdown voltage $V_{(BR)DSS} = 25\,\mathrm{V}$ and single pulse drain-to-source avalanche energy $E = 733\,\mathrm{mJ}$. Standard test conditions $V_{DD} = 20\,\mathrm{V}$, $L = 1\,\mathrm{mH}$, $V_G = 10\,\mathrm{V}$, and $R_G = 25\,\Omega$ were used for measurement and mixed mode electro-thermal simulations. As the behavior of the DMOS transistor under stress is very complex and depends on combined electro-thermal effects, it is necessary to model correctly the experimental device for non-isothermal simulations. While a few μm thick structure is sufficient for electrical simulations, much thicker silicon substrates ($\approx 100\,\mu\mathrm{m}$) must be used for thermal simulations and a tradeoff is a relatively long CPU elapsed time and memory. As the time of the UIS test is very short in ms range, we neglected the thermal conductivity of the package and set the constant boundary temperature $T = 300\,\mathrm{K}$ at the bottom of the structure.

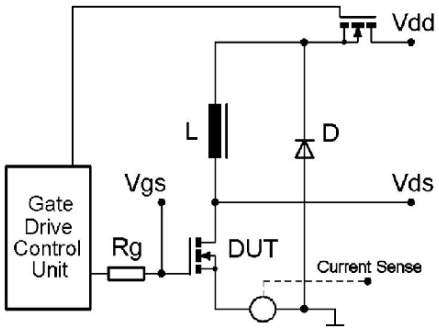


Figure 20. Simple UIS test circuit and corresponding voltage and current waveforms.

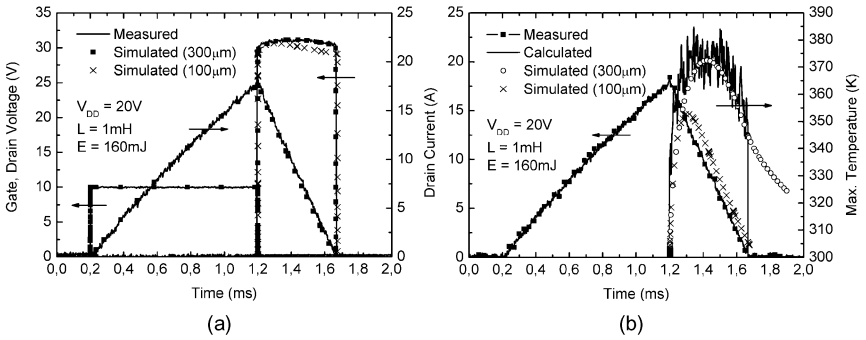


Figure 21. UIS test: (a) measured and simulated waveforms, (b) temperature evolution.

Figure 21 shows the measured and simulated device output current and voltage waveforms and maximum device temperature under UIS test with applied energy of 160 mJ for two structures. They differ in the thickness of the silicon block, the first one had the Si thickness of 100 μm and the second one was 300 μm thick. The experimental device temperature was calculated from the temperature dependence of the static drain-to-source avalanche breakdown voltage $V_{BR(DSS)}$ [31]. Hence, a relatively high noise in the temperature curve can be seen and we have information about the device temperature only during the avalanche regime. However, during the switch-off phase, a high voltage appeared across the device and high current flowed through the device, which caused a great deal of self-heating. It can be seen from the device drain voltage waveform (Figure 21) that the breakdown voltage rises above the starting breakdown voltage value. We can clearly see how important is a proper definition of the geometry of the analyzed structure for non-isothermal simulation. While the simulated electrical characteristics are almost similar for both structures, only a small difference in $V_{(BR)DSS}$ is observable, there is a considerable discrepancy between experimental and simulated maximum temperature dependences with time for the 100 μm thick structure while the agreement for 300 μm structure is excellent.

Figure 22 shows the simulated current, voltage, and temperature waveforms for two different energies during the off state phase of UIS test when the inductor was discharged. For energy $E = 800$ mJ the current related to the reverse biased PN junction at the bottom of the P-well (see Figure 17b) flowed predominantly through the P-well contact, while the current flow through the N-source contact was negligible. However, for energy $E = 1000$ mJ the voltage drop of the drain voltage during the avalanche breakdown can be clearly seen. This voltage drop was caused by opening of the parasitic BJT as indicated by the increased current through N-source and correspondingly decreased current through P-well (Figure 22b). The continuous decrease of the drain current

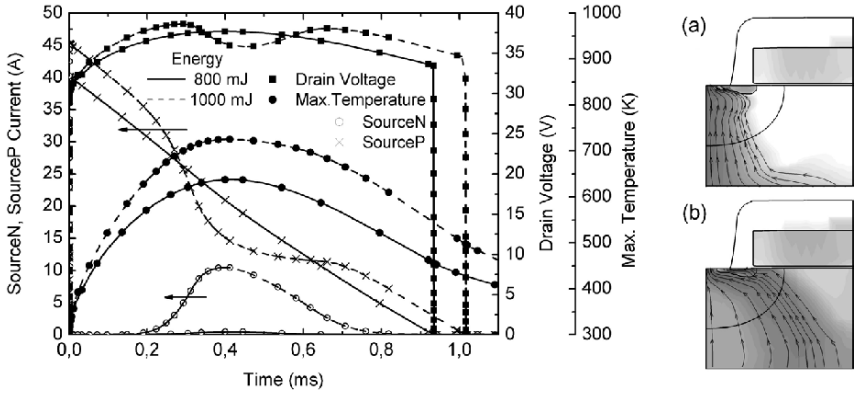


Figure 22. Simulated waveforms of UIS test for different energies (left), detail of the current flow pass in device structure under avalanche breakdown at $t = 0.4$ ms (right): (a) 800 mJ and (b) 1000 mJ.

generated less heat, which resulted in a decrease of the maximum device temperature. Consequently, the parasitic BJT was switched off and the current flowed again through the P-well contact until all energy accumulated in the inductor was dissipated. Numerical simulations with different energies can help to determine the maximum energy which the device can sustain in an ideal case of operation and the behavior and properties of various new structures can be predicted [32, 33].

5. Conclusions

The presented three examples of 2D process and device simulation show how extremely useful tools they are for the analysis, characterization and optimization of fabrication processes and corresponding electro-thermal properties of semiconductor structures and devices. The results of the process and device simulations based on the numerical solution of basic semiconductor equations with complex electro-physical models provide a unique insight into the internal operation of the analyzed devices. Visualization of the internal electrical, thermal, optical, magnetic and mechanical properties allows comprehensive analysis of the critical regions and weak points of the analyzed structures. 2/3D modeling and simulation considerably contribute to a better understanding of the physics of the formation and behavior of parasitic devices that exist as inevitable parts of active devices and degrade their normal operation and reliability. Based on the obtained knowledge, new structures and devices with a modified layout and concentration profiles can be designed and verified.

We report on excellent agreement between the measured and simulated results. Hence, TCAD simulators with properly selected calibrated physical models and defined structures are very fast and cost effective tools for parametric predictive analysis of new technologies, structures and devices integrated in IC's, and also for the physical interpretation of their properties and behavior. The user friendly interactive environment of commercially available TCAD process and device simulators supports their wide use by anybody who is interested in a better understanding of the complex structure and device behavior under various stress conditions.

The key goal of the further development of TCAD tools is to get a time and cost effective vehicle which will provide true simulated results based on more complex physical implemented models, denser structures and/or 3D simulations, and the tradeoff is relatively high CPU time and memory consumption. The problem of getting results with acceptable precision by selection of appropriate models and structures in adequate time must be resolved and optimized for each specific situation.

Acknowledgement

The authors are grateful to Wladek Grabinski for his continuous encouragement in this work. Research on certain topics discussed in this chapter was supported by project APVT 20/0139/02 of the Science and Technology Assistance Agency and grant VEGA 1/2041/05 of the Slovak Ministry of Education.

References

[1] Moore, G.E. "Progress in digital integrated electronics", *IEDM Tech. Digest*, **1975**, 11–13.

[2] Semiconductor Industry Association, International Roadmap for Semiconductors, **2004**, (available at http://public.itrs.net/)

[3] GENESISe-ISE, User manual, ver. 10.0, ISE Zurich, **2004**.

[4] http://www.silvaco.com/products/interactive_tools/vwf.html

[5] Lee, K.; Shur, M.; Ejeldly, T.A.; Ytterdal, T. *Semiconductor Device Modeling for VLSI*. Englewood Cliffs: Prentice Hall, **1993**.

[6] Schenk, A. *Advanced Physical Models for Silicon Device Simulation*. New York: Springer Wien, **1998**.

[7] Jungeman, C.; Meinerzhagen, B. *Hierarchical Device Simulation*. New York: Springer Wien, **2003**.

[8] De Man, H. "Demands on Microelectronics Education and Research in Post – PC Area", *Proceedings of the 3rd EWME*. Dordrecht: Kluwer Academic Publishers, **2000**, 9–14.

[9] Grasser, T.; Jungemann, C.; Kosina, H.; Meinerzhagen, B.; Selberherr, S. "Advanced transport models for sub-micrometer devices", In *Simulation of Semiconductor Processes and Devices*. Wien: Springer-Verlag, **2004**, 1–8.

[10] Rainey, V.P. "Beyond technology – renaissance engineers", *IEEE Trans. Education*, **2002**, *45*, 4–5.
[11] Miura-Mattausch, M.; Mattausch, H.J.; Arora, N.D.; Yang, C.Y. "MOSFET modelling gets physical", *IEEE Circuits and Devices*, **2001**, *17*, 29–36.
[12] *Simulation of Semiconductor Processes and Devices 2004*, G. Wachutka and G. Schrag, Eds. Wien, New York: Springer-Verlag, **2004**.
[13] Zechner, C. *et al*. "New Implantation Tables for B, BF2, P, As, In and Sb", *14th International Conference on Ion Implantation Technology Proceedings*. New Mexico, USA: Taos, **2002**, 567–570.
[14] Ryssel, H.; Krüger, W.; Lorenz, J. "Comparison of Monte-Carlo simulations and analytical models for the calculation of implantation profiles in multilayer targets", *Nucl. Instrum. and Meth.*, **1987**, *B12(20)*, 40–44.
[15] DIOS – ISE, User manual, ver. 10.0, ISE Zurich, **2004**.
[16] Wachutka, G. "An extended thermodynamic model for the simultaneous simulation of the thermal and electrical behavior of semiconductor devices", In *Proc. Sixth Int. NASECODE Conf.*, J.J.H. Miller, Ed., Boole Press Ltd., **1989**, pp. 409–414.
[17] Apanovich, Y.; Lyumkis, E.; Polsky, B.; Shur, A.; Blakey, P. "Steady-state and transient analysis of submicron devices using energy balance and simplified hydrodynamic models", *IEEE Trans. CAD*, **1994**, *13*, 702–710.
[18] DESSIS – ISE, User manual, ver. 10.0, ISE Zurich, **2004**.
[19] Donoval, D.; Chvala, A.; Vrbicky, A. "Computer Aided Analysis of the Parasitic Properties of a Bipolar Transistor Cell", *Proc. 5th EWME*. Dordrecht: Kluwer Academic Publishers, **2004**, 153–158.
[20] Menozzi, *et al*. "Layout dependence of CMOS latch up", *IEEE Trans. Electron Dev.*, **1989**, *36*, 1892–1901.
[21] Ker, M.D.; Lo, W.Y.; Wu, C.Y. "New Experimental Methodology to Extract Compact Layout Rules for Latch up Prevention in Bulk CMOS IC's", *Custom Integrated System Conf.*, **1999**, 143–146.
[22] IC Latch Up Test, EIA/JEDEC Standard No. 78, Electronic Industries Association, **1997**.
[23] Contiero, C.; Andreini, A.; Galbiati, P. "Roadmap Differentiation and Emerging Trends in BCD Technology", *Proc. 32nd Euro. Solid State Research Conference (ESSDERC)*. Italy: Firenze, **2002**, 459–462.
[24] Kawamoto, K.; Takahashi, S.; Fujino, S.; Shirakawa, I. "A no snapback LDMOSFET with automotive ESD endurance", *IEEE Trans. Electron Dev.*, **2002**, *49*, 2047–2053.
[25] Constapel, R.; Shekar, M.S.; Williams, R.K. "Unclamped inductive switching of integrated quasi-vertical DMOSFETs", *IEEE Trans. Electron Dev.*, **1996**, 219–222.
[26] Pinardi, K.; Heinle, U.; Bengtsson, S.; Olsson, J.; Colinge, J.P. "Electrothermal simulations of high-power SOI vertical DMOS transistor with lateral drain contacts under unclamped inductive switching test", *Solid State Electron*, **2004**, *48*, 1119–1126.
[27] Donoval, D.; Vrbicky, A. "Analysis of the Electrical and Thermal Properties of Power DMOS Devices during UIS Supported by 2-D Process and Device Simulation", *Proc. ASDAM*, **2004**, 211–214. Smolenice 2004.
[28] Izaca Deckelmann, A.; Wachutka, G.; Hirler, F.; Krumrey, J.; Henninger, R. "Failure of Multiple-Cell Power DMOS Transistor in Avalanche Operation", *Proc. 33rd European Solid State Res. Conf. (ESSDERC)*. Portugal: Estoril, **2003**, 323–326.
[29] Fischer, K.; Shenai, K. "Dynamics of power MOSFET switching under unclamped inductive loading conditions", *IEEE Trans. Electron Dev.*, **1996**, *43*, 1007–1015.
[30] Fischer, K.; Shenai, K. "Electrothermal effects during unclamped inductive switching (UIS) of power MOSFET's", *IEEE Trans. Electron Dev.*, **1997**, *44*, 874–878.

[31] D'Arcangelo, E. *et. al.* "Experimental characterization of temperature distribution on power MOS devices during unclamped inductive switching", *Microelectronics & Reliability*, **2004**, 1455–1459.

[32] Chien, F. et. al. "High ruggedness power MOSFET design by self-align p^+ process", *IEICE Trans. Electron*, **2005** *E88-C(4)*, 694–698.

[33] Nassif-Khalil, S.G.; Salama, C.A.T. "Super junction LDMOSFET on a silicon-on sapphire substrate", *IEEE Trans. Electron Dev.*, **2003**, *50*, 1385–1391.

Chapter 2

PSP: AN ADVANCED SURFACE-POTENTIAL-BASED MOSFET MODEL

R. van Langevelde[1], G. Gildenblat[2]
[1]*Philips Research Laboratories, High Tech Campus 5, 5656AE Eindhoven, The Netherlands*
E-mail: Ronald.van.Langevelde@philips.com
[2]*Department of Electrical Engineering The Pennsylvania State University University Park, PA 16802 USA*
E-mail: Gildenblat@psu.edu

Abstract: PSP is the latest and the most advanced compact MOSFET model. It was developed by merging and enhancing the best features of the two surface-potential-based models SP (developed at The Pennsylvania State University) and MOS Model 11 (developed by Philips Research). PSP has been selected as a new industry standard for the next generation compact MOSFET model by the Compact Modeling Council. This chapter presents the main ideas enabling the development of PSP, the model structure and its general features.

Key words: compact model; MOSFET; surface potential; PSP, JUNCAP2.

1. Introduction

In computer-aided design of integrated circuits, compact models are used to reproduce electrical characteristics of semiconductor devices. These models describe the device behavior as a function of bias conditions, temperature, device geometry and process variations. For IC-design in CMOS, compact MOSFET models are a critical link in the translation of CMOS process properties into IC performance. In the IC-industry, state-of-the-art compact MOS

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, 29–66.

models in the public domain such as BSIM3 [1], BSIM4 [1] and MOS Model 9 (MM9) [2], are widely used. With the continuous down scaling of CMOS technologies, however, the demands for compact MOS models have become more and more stringent:

- As the supply voltage is scaled down, the moderate inversion region becomes an increasingly larger fraction of the maximum voltage swing. An accurate, physical description of moderate inversion becomes essential, and it can be most easily obtained by the use of surface-potential-based models.
- Modern CMOS technologies are suitable for digital, analog as well as RF applications. The compact model should thus be accurate for digital, analog and RF circuit design. This implies that the model should, amongst others, provide Gummel drain-source symmetry and give an accurate description of distortion behavior.
- The model should accurately describe all the important physical effects of contemporary and future CMOS technologies.

State-of-the-art models such as BSIM4 and MM9 are based on threshold voltage formulations, so-called threshold-voltage-based models, and they fail to fulfil some or all of the above requirements for advanced modeling. This deficiency has presently resulted in a wide consensus in the compact modeling community that traditional threshold-voltage-based models have reached the limit of their usefulness and need to be replaced with more advanced models based on surface potential ψ_s or inversion charge density q_i formulations [3], referred to as surface-potential-based or inversion-charge-based models[1], respectively. The development of the SP model at The Pennsylvania State University [4–17] and MOS Model 11 (MM11) at Philips Research [18–26] has followed the ψ_s-based approach. This approach provides for a physics-based modeling of all regions of operation (including the moderate inversion and the accumulation region) and avoids making additional approximations beyond those already inherent in the charge-sheet models. While the constitutive equation of q_i-based models such as ACM, EKV and BSIM5 [3] can be derived differently, in the final analysis it follows from the equation for surface potential introducing several extra approximations [4]. In addition the ψ_s-based approach, as opposed to the q_i-based approach, enables the physical modeling of the source-drain overlap regions where the inversion charge is not a particularly suitable variable.

The ψ_s-based approach to modeling MOS transistors dates back to the Pao-Sah model [27]. The modern ψ_s-based models are based on the charge-sheet model (CSM) of Brews [28]. Despite the clear physics and the ability to

[1] Here we use the model classification suggested in [4].

provide a single expression for all regions of operation [29] ψ_s-based models did not become popular until the last decade due, in part, to their perceived complexity. Successful ψ_s-based models became possible only after significant progress was made in the techniques for computing the surface potential, simplification of the charge equations relative to the original formulation and the introduction of small-geometry effects. The implementation of these advances and the overall model structures of SP [4] and MM11 [22] turned out to be compatible, enabling the merger of both models into a single new model called PSP that combines and enhances the best features of SP and MM11. This chapter provides an overview of PSP.

The PSP core model contains an intrinsic and an extrinsic model. The intrinsic model describes the electrical behavior of the channel region of the MOSFET, and includes expressions for the drain-source channel current and the quasi-static (QS) terminal charges. The extrinsic model describes the electrical behavior of the gate overlap regions of the MOSFET, and contains expressions for the substrate current, the gate current and the gate overlap and fringing capacitances. PSP also includes a noise model which describes the (intrinsic and extrinsic) noise sources. In addition, PSP provides for two support modules: a new junction model named JUNCAP2 [30] and the non-quasi-static (NQS) module [8, 15, 31].

Both MM11 and SP distinguish between local and global model parameters. This approach is carried over to PSP. Global parameters include geometry dependencies and before evaluating the MOSFET output characteristics they are converted into a small number of local parameters actually used in the core model. The use of local parameters facilitates the model parameter extraction, as one can extract the local parameters for each device geometry separately and then use scaling equations to obtain the global parameters for the relevant range of geometries.

The major features of PSP include the following.

- Physical ψ_s-based formulation of both intrinsic and extrinsic models
- Physical and accurate description of the accumulation region
- Symmetrical linearization enabling accurate modeling of ratio-based circuits (e.g., R2R circuits)
- Gummel symmetry
- Coulomb scattering and non-universality in the mobility model
- Non-singular velocity-field relation enabling the accurate modeling of RF distortion
- Quantum-mechanical corrections
- Correction for polysilicon depletion effects
- Inclusion of all relevant small geometry effects
- modeling of halo implant effects, including the output conductance degradation in long devices

- GIDL/GISL model
- Surface-potential-based noise model including flicker noise, and partly correlated channel thermal noise and channel-induced gate noise.
- Advanced junction model including Shockley-Read-Hall generation/recombination, trap-assisted tunneling and band-to-band tunneling
- Spline-collocation-based NQS model including all terminal currents
- STI-induced stress model

This chapter aims at giving a derivation and physical description of the most important equations used in PSP. Limited space, however, does not allow for discussing all the features included in PSP in detail. For a complete overview of all equations and parameters, the reader is referred to the PSP documentation as can be found on the internet [32]. In Section 2, we will first discuss the intrinsic model, followed by a discussion of the extrinsic model in Section 3. Next, the noise model, the junction diode model and the non-quasi-static model will be treated separately in Sections 4, 5 and 6, respectively. Finally, we will conclude in Section 7.

2. Intrinsic Model

The intrinsic model contains expressions for the drain-source current and the terminal charges. These electrical quantities can be most easily written in terms of the surface potential, hence we start with a discussion of the surface potential in Section 2.1. Next, an approximate method to include two-dimensional effects important for small-geometry devices, the lateral field gradient factor, is treated in Section 2.2. The drain current and the intrinsic charges will be discussed in Sections 2.3 and 2.4, respectively.

2.1. Surface Potential

The surface potential ψ_s is the most natural variable for the formulation of MOS device physics. It is defined as the difference between the electrostatic potential at the SiO_2/Si interface and the potential in the neutral bulk region due to band bending, see Figure 1 (a). Assuming an ideal gate (i.e., neglecting the poly-depletion effect), ψ_s is found using the following derivation [27, 33].

In the p-type substrate, the Poisson equation for the electrostatic potential ψ (with respect to the neutral bulk) is written as:

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} = -\frac{\rho(x,y)}{\epsilon_{\mathrm{Si}}} = q \cdot \frac{N_{\mathrm{SUB}} + n(x,y) - p(x,y)}{\epsilon_{\mathrm{Si}}} \tag{1}$$

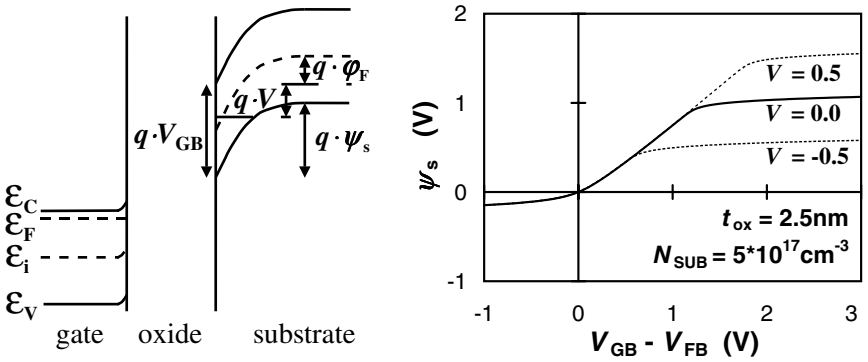


Figure 1. (a) The energy-band diagram (in transversal direction) of an n-MOSFET for $V_{GB} > V_{FB}$, where V_{FB} is the flat-band voltage, ψ_s is the surface potential, V is the difference between electron and hole quasi-Fermi potentials, and ϕ_F is the intrinsic Fermi-potential ($\phi_F = \phi_T \cdot \ln(p_b/n_b)$). (b) The surface potential as a function of gate bias for different values of quasi-Fermi potential V as calculated from (3).

where x and y are the transversal and lateral coordinates, respectively, ρ is the space charge, and N_{SUB} is the net acceptor doping concentration. The electron and hole density, n and p, are given by Maxwell-Boltzmann statistics:

$$\begin{aligned} n(x,y) &= n_b \cdot \exp\left(\frac{\psi(x,y)-V(x)}{\phi_T}\right) \\ p(x,y) &= p_b \cdot \exp\left(-\frac{\psi(x,y)}{\phi_T}\right) \end{aligned} \tag{2}$$

where n_b and p_b denote the electron and hole concentration in the neutral bulk, respectively, ϕ_T ($= k \cdot T/q$) is the thermal voltage, and $V(x)$ denotes the difference between electron and hole quasi-Fermi potentials. This so-called channel voltage $V(x)$ ranges from V_{SB} at the source side ($y = 0$) to V_{DB} at the drain side ($y = L$). Charge neutrality in the bulk sets $N_{\mathrm{SUB}} = p_b - n_b$. In order to obtain an approximate analytical solution of (1), the impact of the lateral field gradient is neglected, i.e., it is assumed that $\partial^2\psi/\partial y^2 \ll \partial^2\psi/\partial x^2$. This is commonly refered to as the gradual channel approximation (GCA). Next, the surface potential ψ_s can be obtained using the first integral of the 1-D Poisson equation and applying Gauss' theorem at the SiO_2/Si interface, where both ψ and $\partial\psi/\partial y$ are taken to be equal to zero deep in the neutral bulk. The resulting equation is the so-called surface potential equation (SPE), which provides ψ_s as an implicit function of the terminal voltage V_{GB} and

the channel voltage V:

$$\left(\frac{V_{GB} - V_{FB} - \psi_s}{\gamma \cdot \sqrt{\phi_T}}\right)^2 = \exp(-u) + u - 1 + \frac{n_b}{p_b} \cdot k_n \cdot \left[\exp(u) - m(u)\right] \tag{3}$$

Here, γ is the body factor given by $\sqrt{2 \cdot q \cdot \epsilon_{\mathrm{Si}} \cdot N_{\mathrm{SUB}}}/C_{ox}$, V_{FB} is the flat-band voltage, $u = \psi_s/\phi_T$, and $k_n = \exp(-V/\phi_T)$. Following the above derivation, the term $m(u)$ is equal to $1 + u/k_n$. Using (3), the surface potential at the source side (ψ_{ss}) and at the drain side (ψ_{sd}) are given implicitly by setting V equal to V_{SB} and V_{DB}, respectively, see Figure 1 (b). It should be pointed out here that the SPE is not only the basis of ψ_s-based models, but also forms the basis of threshold-voltage-based models [33] and inversion-charge-based models [4].

In the SPE, the term $m(u)$ merely affects the $\psi_s(V_{GB}, V)$ dependence in a narrow region near flat band. Nevertheless, the above specific form of $m(u)$ is problematic very near the flat-band voltage where it results in a negative right-hand side of (3) [34]. This has been traced in [14] to the variation of the electron carrier quasi-Fermi potential across the space charge layer[2] neglected in the original formulation [27]. Several different empirical forms of $m(u)$ have been proposed in literature [7, 14, 17, 34] to provide well-conditioned SPE in all regions of operation. In PSP, the expression for $m(u)$ developed for SP-SOI [17] is adopted:

$$m(u) = u + 1 + \frac{u^2}{u^2 + 1} \tag{4}$$

This expression has the following advantages: *(i)* in contrast to [27], it ensures that the right-hand side of (3) is always positive, *(ii)* in contrast to [34], it ensures that $\partial\psi_{ss}/\partial V_{GB} = \partial\psi_{sd}/\partial V_{GB}$ at flat-band allowing one to simply set $\psi_{ss} = \psi_{sd}$ in accumulation without encountering any discontinuities in the derivatives, and *(iii)* in contrast to [7, 14], it is valid even for very negative values of channel voltage ($V < -0.5$V). Eq. (4) produces well-behaved $\psi_s(V_{GB}, V)$ dependence without any differences in the output device characteristics relative to the original formulation. The above modification of the original $m(u)$ does not affect the output device characteristics and is, essentially, invisible to the model user.

Computation of the surface potential as a function of terminal voltages requires the solution of the implicit Eq. (3) and represents a long-standing problem of the MOS device modeling. Almost from the beginning it was addressed both through iterative computations and via analytical approximations[3]. Initially, it was thought that the need for evaluation of the surface potential would negatively affect the model performance. In today's

[2] In other words, the channel voltage V is not only a function of coordinate y but of coordinate x as well.
[3] Look-up tables were used as well.

sophisticated models [4, 25, 38] computation of ψ_s takes only 5–10% of the model execution time and is easily performed using one of several powerful algorithms [4, 25, 32].

The iterative solution of ψ_s was originally pursued in [36], and significantly improved in [25] and [37]. An iterative approach is used efficiently in some of today's surface-potential-based models [3, 25, 37, 38]. On the other hand, the analytical approximation of ψ_s initially pursued in [39] was found to be insufficient for the purpose of transcapacitance modeling (a much more demanding task than modeling of current-voltage characteristics [40]) and abandoned. This approach – based on obtaining the asymptotic approximations of the surface potential in different regions of MOSFET operation and joining them via smoothing functions – has been further developed in [41] and brought into its most successful form with about 1mV accuracy in MM11 [18] (where it was later replaced by iterative calculations [25]).

A different approach in which the surface potential is obtained by an approximate solution of the SPE was developed in [5, 4, 16]. The analytical approximation in [4] is based on a specific form of $m(u)$ [7, 14] and as such is limited to $V > -0.5\,\mathrm{V}$. In PSP we use an even more powerful analytical approximation based on (4) [17, 32] which is accurate under all bias conditions. Typical results are shown in Figure 2 for both positive and negative bias on the source-drain *pn* junction. The accuracy of this approximation is better than 1nV, which is sufficient for even the most demanding MOSFET modeling

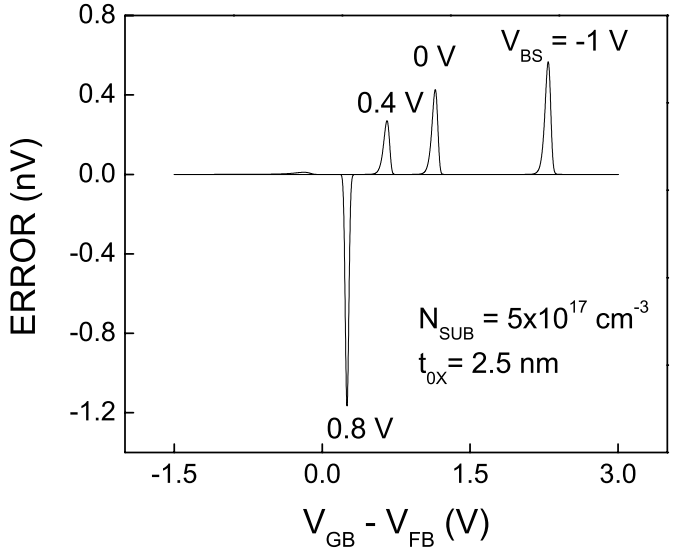


Figure 2. Absolute error of the analytical approximation for the surface potential at source side ψ_{ss} for different values of bulk-source bias V_{BS}.

applications. The approximations of this type are slightly slower than the more simple approximation in [18], but the overall effect on the model execution time is minimal (about 0.4%).

In this section we have discussed the computation of the surface potential in the active region of the device. In the source-drain overlap regions the problem is even simpler and it is addressed in Section 3.1.

2.2. Lateral Field Gradient Factor

As discussed in the previous section, the derivation of the SPE is based on the gradual channel approximation. This approximation neglects the lateral field gradient, and as a result, short-channel effects such as drain-induced barrier lowering (DIBL) and threshold-voltage roll-off are not accurately incorporated in any model based on the GCA.

To extend the model formulation beyond the gradual channel approximation PSP relies on the lateral field gradient factor f introduced in [42]. In weak inversion where $n, p \ll N_{\mathrm{SUB}}$, Eq. (1) at the SiO_2/Si interface is rewritten to:

$$\frac{\partial^2 \psi_s}{\partial x^2} = \frac{q \cdot N_{\mathrm{SUB}}}{\epsilon_{\mathrm{Si}}} \cdot \left(1 - \frac{\epsilon_{\mathrm{Si}}}{q \cdot N_{\mathrm{SUB}}} \cdot \frac{\partial^2 \psi_s}{\partial y^2}\right) = \frac{q \cdot N_{\mathrm{SUB}}}{\epsilon_{\mathrm{Si}}} \cdot f \tag{5}$$

The use of factor f allows the introduction of an effective doping concentration $N_{\mathrm{SUB}} \cdot f$. The application of this method to threshold voltage was reported in [43]. The initial application f this method to surface potential used the bias-independent approximation $f = f(L, W)$ [44], but in PSP, as in SP [4], a bias-dependent approximation is used for f.

An elementary expression for f can be obtained by the following generalization of the analysis in [43, 44]. A parabolic dependence of $\psi_s(y)$ is assumed, which is equivalent to a position-independent f. The boundary conditions are $\psi_s(0) = V_{SB} + V_{BI}$ and $\psi_s(L) = \psi_s(0) + V_{DS}$, where V_{BI} is the built-in potential of the n^+/p source-bulk and drain-bulk junctions. Linearizing the result, one finds the generic expression:

$$f = F_0 \cdot \left(1 - A_f \cdot V_{SB} - C_f \cdot V_{DS}\right) + B_f \cdot \psi_f = f_0 + B_f \cdot \psi_f \tag{6}$$

where ψ_f is the surface potential without lateral field gradient, and F_0, A_f, B_f and C_f are geometry-dependent factors. Despite its simplicity Eq. (6) contains the essential physics: a linear dependence of f on the surface potential ψ_f and a decrease of f with V_{SB} and V_{DS}. The latter can be effectively regarded as the drain-induced barrier lowering (DIBL) effect.

The above derivation serves as a motivation for the actual expression for the lateral gradient factor used in PSP. While the linear dependence of $f(\psi_f)$

has been retained, the dependence on V_{SB} and V_{DS} has been modified in order to assure $f_0 > 0$ for all terminal biases. The result is still Eq. (6) but with

$$f_0 = \frac{F_0}{1 + F_{SB}(V_{SB}) + F_{\mathrm{DIBL}}(V_{DS})} \tag{7}$$

where F_0 is a geometry dependent factor and functions F_{SB}, and F_{DIBL}, as well as the surface potential ψ_f, are selected in a manner consistent with the Gummel symmetry of the model. Complete expressions and further details can be found in [32].

2.3. Drain Current

An important objective of the PSP project is to incorporate essential device physics without a prohibitive increase in the model complexity in the framework of ψ_s-based models. To a large extent this is accomplished using the symmetric linearization technique developed in [4, 6] and similarly in [20, 24]. To simplify the exposition of this key idea we start by reformulating Brews' charge-sheet model (CSM) [28], while neglecting all short-channel effects (which, of course, are included in the complete PSP model equations, see below). There are several ways to arrive at the CSM equations. A particularly simple derivation [29] starts with equation

$$I_{DS} = \mu \cdot W \cdot \left(q_i \cdot \frac{\mathrm{d}\psi_s}{\mathrm{d}y} - \phi_T \cdot \frac{\mathrm{d}q_i}{\mathrm{d}y} \right) \tag{8}$$

where μ denotes the effective channel mobility and q_i is the inversion charge per unit area. There are numerous issues that need to be discussed in connection with the validity of this equation. References [35, 46] and those cited therein are quite useful in this regard. The bottom line is that (8) leads to the original CSM [28] that is justified by comparison with the Pao-Sah model [27]. Our task here is to further simplify (8) in order to make it conducive to the development of a compact MOSFET model.

The symmetric linearization method is based on the approximation

$$q_i = q_{im} - \alpha \cdot (\psi_s - \psi_m) \tag{9}$$

where q_{im} is the inversion charge density at the potential midpoint $\psi_s = \psi_m$:

$$\psi_m = \frac{\psi_{ss} + \psi_{sd}}{2} \tag{10}$$

In the above α denotes the linearization coefficient easily obtained using standard CSM equations [4, 6]. In the full PSP equations the expression for α is slightly more complex in order to provide smooth behavior in all modes of

operations including the region $\psi_s < 3 \cdot \phi_T$ where equations of the original CSM model do not apply. Using (9), Eq. (8) reduces to

$$I_{DS} = \mu \cdot W \cdot q_i^* \cdot \frac{\mathrm{d}\psi_s}{\mathrm{d}y} \tag{11}$$

where q_i^* is the effective inversion charge density modified to account for the diffusion current component

$$q_i^* = q_i + \alpha \cdot \phi_T \tag{12}$$

Integrating from source to drain yields [6]

$$I_{DS} = \mu \cdot \frac{W}{L} \cdot q_{im}^* \cdot \Delta\psi \tag{13}$$

where q_{im}^* is the effective inversion charge density at the surface potential midpoint ψ_m, and $\Delta\psi$ is given by:

$$\Delta\psi = \psi_{sd} - \psi_{ss} \tag{14}$$

The above equation for the drain current is numerically equivalent to the one in the original CSM [28] but is significantly simpler. In particular, fractional powers that are present in the drain current expression in [28] are eliminated, while both drift and diffusion components of the drain current are retained and simplified. Typical results are shown in Figures 3 and 4 indicating that the difference between (13) and the original CSM is less than 1–2% and is inconsequential for the purpose of compact modeling.

Note that Eq. (13) is also accurate in the subthreshold region. In this operation region where $q_{im} \ll \alpha \cdot \phi_T$ and $\Delta\psi$ is an exponential function of the gate bias [9] one can easily recover the classic subthreshold approximation [28, 29, 33].

Up till this point in the derivation of drain current, the carrier mobility in the inversion layer has been assumed constant. In reality, however, this is not true. Carriers in the channel undergo increased scattering with increasing fields, when they move under the influence of the normal electric field and the lateral electric field due to the gate bias V_{GS} and the drain bias V_{DS}, respectively. The former is referred to as mobility reduction, whereas the latter is referred to as velocity saturation.

Mobility Reduction: In a MOS structure the normal electric field restricts the channel to a sheet layer in which two-dimensional confinement effects and scattering cause the mobility to depend on bias conditions. Mobile carriers in the inversion layer can be scattered by ionized doping atoms (so-called Coulomb scattering), by vibrations of the crystal lattice (so-called phonon scattering) and

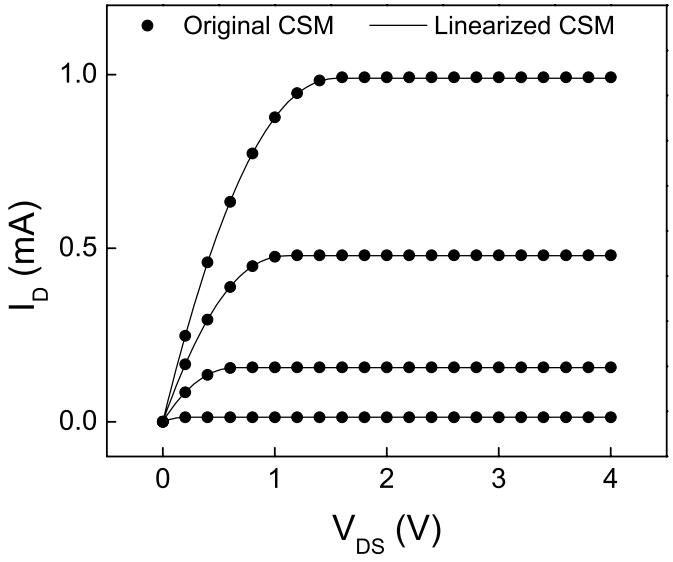


Figure 3. Comparison between the symmetrically linearized and original charge-sheet model; $N_{\mathrm{SUB}} = 5 \cdot 10^{23}\,\mathrm{m}^{-3}$, $t_{ox} = 2\,\mathrm{nm}$, $V_{BS} = 0\,\mathrm{V}$, $\mu = 5 \cdot 10^{-2}\,\mathrm{m}^2/\mathrm{Vs}$, $W/L = 1$, $V_{FB} = -0.9\,\mathrm{V}$, V_{GS} varies between 0.5 and 2 V with 0.5 V steps.

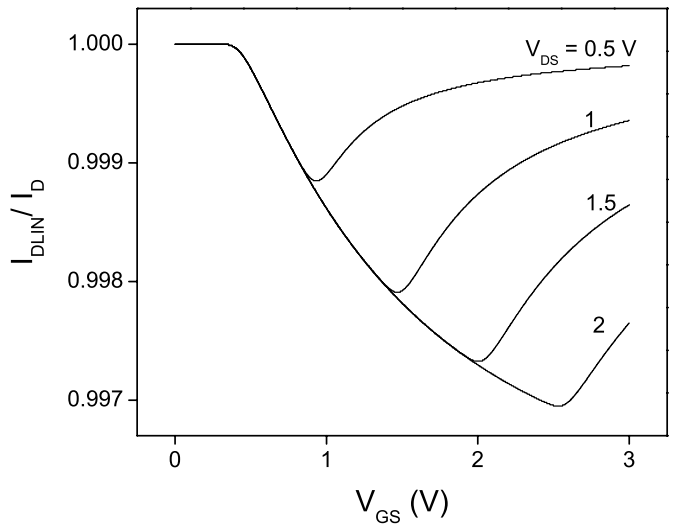


Figure 4. Ratio of the drain currents in symmetrically linearized (I_{DLIN}) and original (I_D) charge-sheet model; $N_{\mathrm{SUB}} = 5 \cdot 10^{23}\,\mathrm{m}^{-3}$, $t_{ox} = 2\,\mathrm{nm}$, $V_{BS} = 0\,\mathrm{V}$, $V_{FB} = -0.8\,\mathrm{V}$.

by the SiO_2/Si interface roughness (so-called surface roughness scattering). The mobility expression used in PSP takes all this into account and is given by:

$$\mu = \mu_{\text{eff}} = \frac{\text{MU0} \cdot \mu_x}{1 + (\text{MUE} \cdot E_{\text{eff}})^{\text{THEMU}} + \text{CS} \cdot \left(\dfrac{q_{bm}}{q_{bm} + q_{im}}\right)^2 + G_R} \tag{15}$$

where MU0 is the low-field mobility, and parameters MUE and THEMU account for the mobility degradation caused by the surface roughness and phonon scattering by the effective vertical field E_{eff}:

$$E_{\text{eff}} = \frac{q_{bm} + \eta \cdot q_{im}}{\epsilon_{\text{Si}}} \tag{16}$$

with $\eta = 1/2$ for electrons and $\eta = 1/3$ for holes. Coulomb scattering is introduced as in [47] using parameter CS, q_{bm} is the bulk charge per unit channel area at the surface potential midpoint [4] and the factor μ_x describes non-universality effects and also accounts (empirically) for doping non-uniformity. The term G_R accounts for the series resistance:

$$G_R = \text{MU0} \cdot \frac{W}{L} \cdot q_{im} \cdot \text{RS} \tag{17}$$

where RS is the source/drain series resistance. When series resistance is included externally G_R can be set to zero.

Velocity Saturation: With an increase in lateral electric field, carriers gain sufficient energy to be scattered by optical phonons, resulting in a decrease of mobility and eventually resulting in the saturation of drift velocity. Velocity saturation is critical not only for the accurate modeling of the saturation region, but also to ensure nonsingular behavior of the model at zero drain bias [45, 48]. The saturation velocity model used in PSP is that of MM11 [22], which is based on the Scharfetter-Gummel expression [49]. For n-channel devices:

$$v_d = \frac{\mu_{\text{eff}} \cdot E_y}{\sqrt{1 + \left(\dfrac{\mu_{\text{eff}}}{v_{\text{sat}}} \cdot E_y\right)^2}} \tag{18}$$

where E_y is the lateral component of the electric field and v_{sat} denotes the saturation velocity. Using (18) in the derivation of drain current leads to an implicit expression for I_{DS}, linearizing this expression leads to the following explicit expression [25]:

$$I_{DS} = \mu_{\text{eff}} \cdot \frac{W}{L} \cdot \frac{q_i^* \cdot \Delta\psi}{G_{\text{vsat}}} \tag{19}$$

where $\theta_{\text{sat}} = \mu_{\text{eff}}/(v_{\text{sat}} \cdot L)$ and:

$$G_{\text{vsat}} = \frac{1}{2} + \frac{1}{2} \cdot \sqrt{1 + 2 \cdot (\theta_{\text{sat}} \cdot \Delta\psi)^2} \tag{20}$$

For p-channel devices, the velocity saturation is accurately described by [49]:

$$v_d = \frac{\mu_{\text{eff}} \cdot E_y}{\sqrt{1 + \dfrac{\left(\mu_{\text{eff}} \cdot E_y/v_c\right)^2}{G + \mu_{\text{eff}} \cdot E_y/v_c}}} \tag{21}$$

where v_c is a parameter corresponding to the velocity of the longitudinal acoustic phonons and G is a fitting parameter. In this case, the integration along the channel is less straightforward. For simplicity's sake, we approximate the term $G + \mu_{\text{eff}} \cdot E_y/v_c$ by $G + \theta_{\text{sat}} \cdot \Delta\psi$ where $\theta_{\text{sat}} = \mu_{\text{eff}}/(v_c \cdot L)$. The parameter G has been found to be of minor influence, and is set equal to 1. In other words, all equations derived for n-channel devices can simply be re-used for p-channel devices by replacing θ_{sat} by $\theta_{\text{sat}}/\sqrt{1 + \theta_{\text{sat}} \cdot \Delta\psi}$.

The resulting expressions for n- and p-channel devices are non-singular, enabling for example the modeling of passive RF mixers [48]. As shown in [19] they also enable accurate modeling of RF distortion in the saturation region.

Long-channel surface-potential-based models automatically include the pinch-off behavior in the saturation region. Pinch-off implies that the channel at the drain end is forced into weak inversion and that the mobile charge density at the drain approaches zero. In reality, however, the description of pinch-off is not realistic, since carriers reach velocity saturation at the drain end before the pinch-off condition is fulfilled. As a result the drain-source saturation voltage V_{dsat} may differ significantly from the pinch-off voltage, and this difference needs to be taken into account in the model. This is a general problem for any compact MOSFET model based on the gradual channel approximation.

In PSP, the saturation voltage V_{dsat} is calculated from setting $\partial I_{DS}/\partial\Delta\psi = 0$. Next, the drain-source voltage V_{DS} is replaced by an effective drain-source voltage V_{dse}, which changes smoothly from V_{DS} in the linear region (i.e., for $V_{DS} \ll V_{\text{dsat}}$) to V_{dsat} in the saturation region (i.e., for $V_{DS} \geq V_{\text{dsat}}$). The smooth transition is obtained by [45]:

$$V_{\text{dse}} = \frac{V_{DS}}{\left[1 + (V_{DS}/V_{\text{dsat}})^{a_x}\right]^{1/a_x}} \tag{22}$$

where a_x (≥ 2) is a local parameter which determines the smoothness of the transition. The use of (22) ensures preservation of Gummel drain-source symmetry [45].

For an accurate description of output conductance $g_{DS} = \partial I_D/\partial V_{DS}$ PSP also includes detailed description of channel length modulation. This description is based on [50] and has been extended to include the impact of pocket implants similar to [51].

The incorporation of mobility reduction, velocity saturation, saturation voltage and channel length modulation as described above results in an accurate description of the output characteristics as shown in Figure 5. In addition, the linearization scheme adopted in PSP (as well as those in SP and MM11) enables accurate modeling of ratio-based circuits. A detailed discussion including applications to R2R circuits can be found in [26].

2.4. Intrinsic Charges

In a quasi-static approximation, charges can be attributed to the four terminals of the MOSFET: Q_G, Q_D, Q_S and Q_B. Using these charges, one can define 16 transcapacitances C_{ij} (9 of which are independent):

$$C_{ij} = \begin{cases} \dfrac{\partial q_i}{\partial V_j} & \text{for: } i = j \\ -\dfrac{\partial q_i}{\partial V_j} & \text{for: } i \neq j \end{cases} \tag{23}$$

where i and j denote the terminal S, D, G or B. The total gate charge Q_G is calculated by integrating the gate charge density q_g along the channel:

$$Q_G = W \cdot \int_0^L q_g \cdot \mathrm{d}y \tag{24}$$

where $q_g = q_i + q_b = C_{ox} \cdot (V_{GB} - V_{FB} - \psi_s)$. Note that q_g is a simple function of ψ_s, and as a result the calculation of Q_G is quite straight-forward in ψ_s-based models. In threshold-voltage-based and inversion-charge-based models, on the other hand, the surface potential is not readily available and the calculation of Q_G is more elaborate.

The total inversion-layer charge is split up into a source Q_S and a drain Q_D charge. For MOSFETs with a homogeneous doping concentration the Ward-Dutton charge partitioning scheme [52] is valid, and Q_S and Q_D are given by:

$$Q_S = -W \cdot \int_0^L (1 - y/L) \cdot q_i \cdot \mathrm{d}y \tag{25}$$

$$Q_D = -W \cdot \int_0^L y/L \cdot q_i \cdot \mathrm{d}y \tag{26}$$

This partitioning scheme results in bias-dependent or dynamic charge partitioning. Finally, since charge neutrality holds for the complete transistor, the total bulk charge Q_B is simply given by $-Q_S - Q_D - Q_G$.

Since inversion charge q_i and gate charge q_g are functions of the surface potential ψ_s, calculation of these integrals requires $y\,(\psi_s)$ dependence. For the charge-sheet model explicit expressions for the terminal charges have been

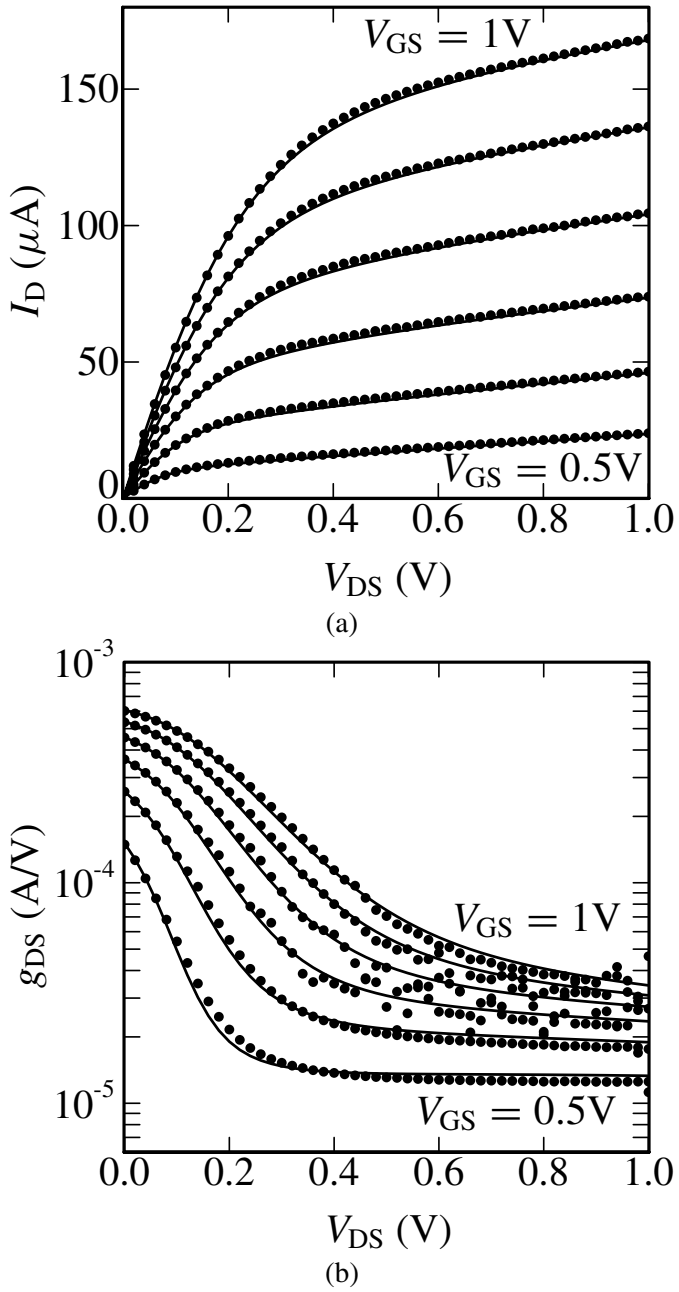


Figure 5. Drain current I_D (a) and corresponding conductance g_{DS} (b) versus drain-source bias V_{DS} for a $W/L = 360\,\text{nm}/90\,\text{nm}$ n-channel MOSFET; V_{GS} varies between 0.5 and 1 V and $V_{SB} = 0\,\text{V}$. Symbols denote measurements and lines represent modeled results using PSP.

given in [34] and, subsequently, in an equivalent but less singular form in [8]. These equations are extremely complex and hence unsuitable for compact modeling purposes. However, just as in the case of the drain current, the symmetric linearization method allows one to derive extremely simple yet accurate expressions numerically indistinguishable from the expressions given in [8, 34]. To simplify the exposition and verification of the technique we first consider the long-channel case and later indicate how the resulting equations can be modified to account for velocity saturation.

From Eqs. (9) through (12), we find:

$$\frac{\mathrm{d}y}{\mathrm{d}s} = \frac{\mu \cdot W}{I_{DS}} \cdot (H - s) \tag{27}$$

where $s = \psi_s - \psi_m$ and $H = q^*_{im}/\alpha$. Separating variables and integrating, we finds [6, 8]:

$$\psi_s(y) = \psi_m + H \cdot \left[1 - \sqrt{1 - 2 \cdot \frac{\Delta\psi}{H} \cdot \frac{y - y_m}{L}}\right] \tag{28}$$

where y_m denotes the coordinate of the surface potential midpoint ψ_m:

$$y_m = \frac{L}{2} \cdot \left(1 + \frac{\Delta\psi}{4 \cdot H}\right) \tag{29}$$

This result of the symmetric linearization method can be compared with the $y\,(\psi_s)$ dependence obtained from the charge-sheet model. Typical plots shown in Figure 6 and given in [6, 8] indicate the high accuracy of (27) and (28).

With (27) available it is a simple matter to compute the integrals for the terminal charges by changing variables from y to s. For example, Eq. (26) for Q_D results in:

$$Q_D = \frac{q_{im}}{2} + \alpha \cdot \frac{\Delta\psi}{12} \cdot \left(1 - \frac{\Delta\psi}{2 \cdot H} - \frac{\Delta\psi^2}{20 \cdot H^2}\right) \tag{30}$$

To verify the accuracy of expression (30) and similar expressions for other terminal charges, they are compared with the exact results in [8, 34]. To make this comparison particularly stringent we evaluate the transcapacitances C_{ij}.

The results shown in Figure 7 indicate that symmetric linearization is extremely accurate. Two comments can be made concerning this conclusion. Firstly, the integration along the channel is a common task in the development of a compact MOSFET model. It is involved in the evaluation of the gate current, the noise spectral densities, etc. In all cases symmetric linearization allows one to obtain manageable equations without compromising the device physics. Secondly, all compact models (even the older threshold-voltage based ones [33, 53]) include some form of linearization of the inversion charge as a function of the surface potential in order to escape complicated expressions

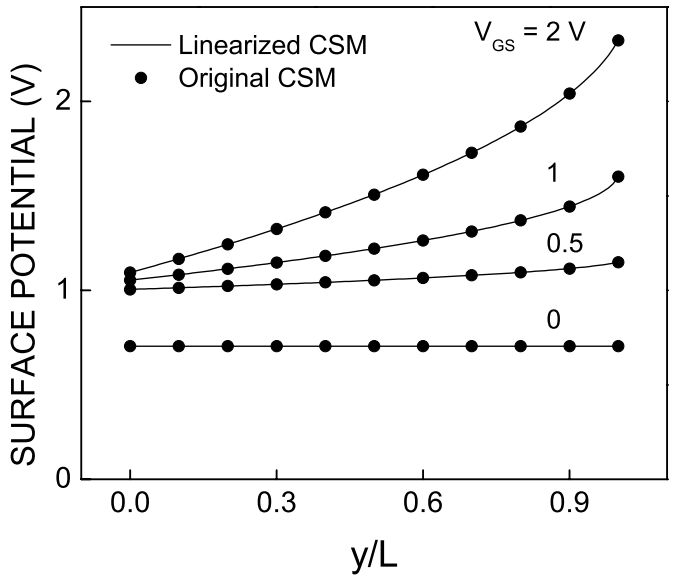


Figure 6. Comparison of the position dependence of surface potential for symmetrically linearized and original charge-sheet models; $N_{\mathrm{SUB}} = 5 \cdot 10^{23}\,\mathrm{m}^{-3}$, $t_{ox} = 2\,\mathrm{nm}$, $V_{BS} = 0\,\mathrm{V}$, $V_{FB} = -0.9\,\mathrm{V}$.

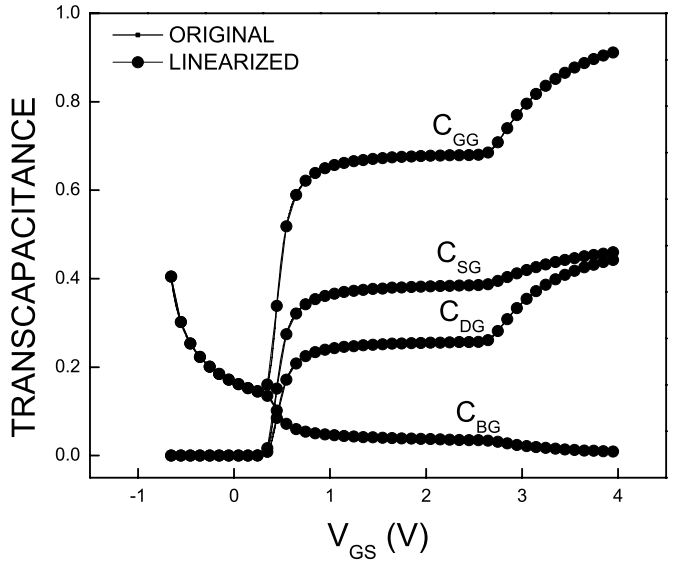


Figure 7. Comparison of transcapacitances for linearized and original charge-sheet models; $N_{\mathrm{SUB}} = 5 \cdot 10^{23}\,\mathrm{m}^{-3}$, $t_{ox} = 2.5\,\mathrm{nm}$, $V_{FB} = -0.8\,\mathrm{V}$, $V_{BS} = 0\,\mathrm{V}$, $V_{DS} = 2\,\mathrm{V}$.

for the terminal charges. In many of the traditional models this results in the loss of Gummel symmetry [26, 33, 45, 48]. In addition, the complexity of the charge expressions may necessitate decoupling the charge and the current expressions with the well-known unfortunate consequences for circuit simulations described, e.g., in [54]. The symmetric linearization method solves both of these problems without complicating the model structure. In fact, the resulting expressions are simpler than in the traditional approach.

The key to the merger of SP and MM11 is the inclusion of the different expression for the drift velocity (18) and the drain current (19) within the context of symmetric linearization. The initial version of this technique was developed for long-channel devices to verify the concept. It was later shown that the flexibility of the symmetric linearization method is such that Eqs. (27) through (30) remain unchanged when the velocity saturation model in SP is included; the only difference being the change in the expression for H [4]. This approach is carried over to PSP where the position dependence of ψ_s is still given by (28), but in order to accommodate the different expression for the drift velocity and the drain current, it can be derived that:

$$H_{\mathrm{PSP}} = \frac{q_i^*}{\alpha' \cdot G_{\mathrm{vsat}}} \tag{31}$$

where

$$\alpha' = \alpha \cdot \left[1 + \frac{1}{2} \cdot \left(\frac{\theta_{\mathrm{sat}} \cdot \Delta\psi}{G_{\mathrm{vsat}}}\right)^2\right] \tag{32}$$

With this in mind the quasi-static terminal charges can be evaluated as in [4, 6, 8], the only difference being that now $H = H_{\mathrm{PSP}}$. For example, the normalized drain charge given in the Ward-Dutton partition is still given by (30). The expressions for the current and terminal charges obtained in this manner are continuous and smooth in all regions of operation from accumulation to strong inversion.

3. Extrinsic Model

The extrinsic model includes contributions of the gate/source and gate/drain overlap regions, and the gate and bulk current. As is the case for the intrinsic model, the electrical behavior in the overlap regions can be most easily described in terms of the surface potential. Consequently, we will start with a discussion of the surface potential in the overlap regions in Section 3.1. Next, the bulk current will be discussed in Section 3.2, followed by a discussion of gate current in 3.3. Finally, the extrinsic charges and capacitances will be treated in Section 3.4.

3.1. Surface Potential in the Overlap Regions

For a quantitative description of the gate/source and gate/drain overlap regions, the overlap regions are treated as n^+-gate/oxide/n^+-bulk MOS capacitances, where the source (or drain) acts as bulk terminal. Assuming the doping profile in the n^+-source extension can be approximated by a uniform constant doping concentration N_{OV}, we can define a body factor γ_{ov} and a flat-band voltage V_{FBov} in this region. A surface potential ψ_{ov} can be calculated (both at source and drain side) using the SPE (3), which can be further simplified by neglecting the minority carrier contribution to the space charge[4]:

$$\left(\frac{V_{GX} - V_{\mathrm{FBov}} - \psi_{ov}}{\gamma_{ov} \cdot \sqrt{\phi_T}}\right)^2 = \exp(-u_{ov}) + u_{ov} - 1 \tag{33}$$

where $u_{ov} = \psi_{ov}/\phi_T$ and V_{GX} denotes either V_{GS} or V_{GD}. Note that to facilitate the comparison with (3), Eq. (33) is written for the p^+ overlap region, i.e., for the case of p-channel transistors. In n-channel devices with n^+ overlap regions one needs to make obvious sign changes in (33).

Analytical approximation for the non-iterative solution of this equation has been initially given in [12] and the final version can be found in [32]. Typical results are shown in Figure 8 for the cases of high and moderate doping, respectively. While the high doping levels are more important for the modeling of the overlap regions, this analytical approximation appears (in a totally different physical context) in the problem of dynamic varactor modeling [55] and in the development of the non-quasi-static model [15]. Hence, it is essential that the accuracy of the approximation is quite high regardless of the doping level.

The derivation of currents and charges in the overlap regions is most easily performed in terms of the oxide voltage in the overlap region V_{ov}, which is simply given by:

$$V_{ov} = V_{GX} - V_{\mathrm{FBov}} - \psi_{ov} \tag{34}$$

This quantity is extensively used in the following sections on bulk current, gate current and extrinsic charges.

3.2. Bulk Current

Up to this point, it has been assumed that the bulk current in a MOSFET is equal to zero. Bulk current may, however, be generated between drain and bulk or between source and bulk by impact ionization and gate-induced drain

[4]This approach disallows description of the inversion channel but since the source/drain extension is highly doped, the inversion channel can only be formed at unrealistically negative gate-source or gate-drain bias.

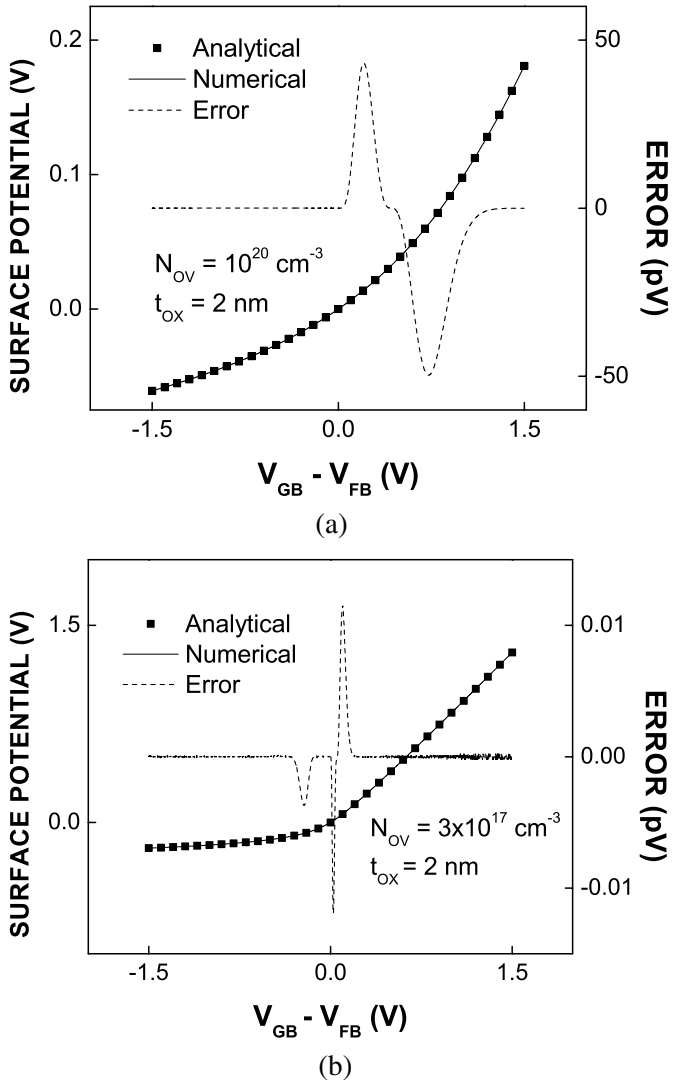


Figure 8. Absolute error of the analytical approximation for the surface potential neglecting the minority carrier contribution in (a) a highly doped the source/drain overlap region and (b) a moderately doped region.

leakage (GIDL). These effects are all included in PSP and are briefly discussed in this section.

Impact Ionization: Subjected to a high lateral electric field, electrons in the channel will accelerate traveling from source to drain and gain so much energy that they can create extra electron-hole pairs by exciting electrons from the

valence band into the conduction band. This effect is generally referred to as impact ionization, and it results in a current I_{ii} between drain and bulk. The impact-ionization current is conventionally written as [33]:

$$I_{ii} \propto I_{DS} \cdot E_m \cdot \exp\left(-b \,/\, E_m\right) \tag{35}$$

where b is a parameter and E_m is the maximum lateral field in the channel. In PSP, this conventional description has been extended with an accurate description of the subthreshold region and the impact of back bias [9].

Gate-Induced Drain Leakage: When the MOSFET is in off-state, a significant leakage current flowing from drain to bulk can be detected at a drain voltage much lower than the breakdown voltage [56]. This drain leakage current is caused by the gate-induced high electric field in the gate-to-drain overlap region, and as a result it has been named gate-induced drain leakage (GIDL). For negative gate-drain bias V_{GD}, a high transversal field is created in the depletion region formed in the gate-to-drain overlap region. Electron-hole pairs are generated by the band-to-band tunneling[5] of valence band electrons into the conduction band and collected by the drain and bulk separately. A simple expression for GIDL current based on [57] is given by:

$$J_{\mathrm{GIDL}} \propto {E_{\mathrm{tov}}}^2 \cdot \exp\left(-B^*_{\mathrm{GIDL}} / E_{\mathrm{tov}}\right) \tag{36}$$

where B^*_{GIDL} is a physical parameter and E_{tov} is the maximum electric field at the Si/SiO_2-interface in the drain overlap region. The latter consists of a (dominant) transversal component (equal to $C_{ox} \cdot V_{ov}/\epsilon_{\mathrm{Si}}$) and a lateral component empirically proportional to V_{DB}. The maximum electric field E_{tov} can be written as:

$$E_{\mathrm{tov}} = \frac{C_{ox}}{\epsilon_{\mathrm{Si}}} \cdot \sqrt{{V_{ov}}^2 + (C_{\mathrm{GIDL}} \cdot V_{DB})^2} = \frac{C_{ox}}{\epsilon_{\mathrm{Si}}} \cdot V_{\mathrm{tov}} \tag{37}$$

where C_{GIDL} is an empirical parameter. Using (36) and (37), we can write for the total GIDL current:

$$I_{\mathrm{GIDL}} = A_{\mathrm{GIDL}} \cdot V_{DB} \cdot {V_{\mathrm{tov}}}^2 \cdot \exp\left(-B_{\mathrm{GIDL}} / V_{\mathrm{tov}}\right) \tag{38}$$

where $A_{\mathrm{GIDL}} \propto W \cdot \Delta L_{ov} \cdot C_{ox}/\epsilon_{\mathrm{Si}}$ and $B_{\mathrm{GIDL}} = \epsilon_{\mathrm{Si}} \cdot B^*_{\mathrm{GIDL}}/C_{ox}$, but they are both considered as local parameters. The V_{DB} term in (38) is empirical and has been added in order to ensure that $I_{\mathrm{GIDL}} = 0$ for $V_{DB} = 0$ and that I_{GIDL} changes sign when V_{DB} changes sign.

In the above derivation we have focussed on the gate-induced drain leakage. The same phenomenon, however, can also occur at the source side, in which case it is referred to as gate-induced source leakage (GISL). The electric field

[5] Trap-assisted tunneling may also occur, but it is neglected in the calculation of GIDL.

in the overlapped source region is typically not as high as the field in the drain region, and as a result, GISL will not really impact the source leakage. Nonetheless, GISL has been incorporated in the PSP model in order to preserve Gummel drain-source symmetry.

3.3. Gate Current

From a classical point of view, gate current in a MOSFET is non-existent, since carriers in the inversion layer cannot cross the potential barrier χ_B of the gate oxide, see Figure 9 (where $\chi_B = \chi_{B_N}$ for electrons and $\chi_B = \chi_{B_P}$ for holes). From a quantum-mechanical point of view, however, carriers may tunnel through the potential barrier resulting in a non-zero gate current density J_G. The probability of tunneling increases exponentially with decreasing oxide thickness t_{ox}, resulting in an exponentially increasing J_G. With CMOS technology scaling, t_{ox} is continuously scaled down, and consequently gate current can no longer be neglected for modern and future CMOS technologies as it may start to affect circuit performance [60, 61]. PSP provides for a gate current

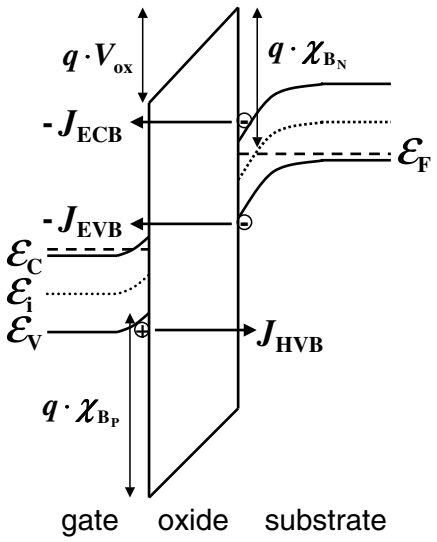


Figure 9. Energy-band diagram of an n-MOS in inversion where χ_{B_N} and χ_{B_P} are the oxide potential barriers for electrons and holes, respectively. Carriers may tunnel through the gate oxide resulting in a non-zero gate current density J_G. Three major mechanisms of gate tunneling can be distinguished: electron conduction-band tunneling (J_{ECB}), electron valence-band tunneling (J_{EVB}) and hole valence-band tunneling (J_{HVB}). ECB tunneling is important for n-MOS devices, whereas HVB tunneling is important for p-MOS devices. EVB tunneling only becomes important for high V_{ox}, and is therefore neglected in the remainder of this section.

model that accurately describes gate leakage in MOSFETs. This gate current model is a further development of the gate current model in SP [12], which in itself is an extension of the gate current model in MM11 [21].

In a typical MOSFET structure, we can distinguish two main gate current components: the gate-to-channel I_{GC} and the gate overlap component I_{Gov}. In the channel or overlap regions of an n-type MOSFET, mainly conduction band tunneling (ECB) is important[6]. The gate current density J_G due to direct tunneling is written as [12]:

$$J_G(y) = J_0 \cdot F_S(y) \cdot D(y) \tag{39}$$

where J_0 is a physical constant, $F_S(y)$ is the supply function [62] and $D(y)$ is the tunneling transmission coefficient. Based on the WKB approximation $D(y)$ is given by:

$$D(y) = \exp\left[-B \cdot f(z_g)\right] \tag{40}$$

where B is a physical constant, z_g is equal to V_{ox}/χ_B, $V_{ox} = q_g/C_{ox}$, and:

$$f(z_g) = \frac{1 - (1 - z_g)^{3/2}}{z_g} \approx -\frac{3}{2} + G_2 \cdot z_g + G_3 \cdot z_g^2 \tag{41}$$

Ideally, the coefficients $G_2 = 3/8$ and $G_3 = 1/16$ can be obtained from a second-order Taylor expansion. However, here they have been turned into adjustable parameters to absorb inaccuracies included in the derivation of (39)–(40). The supply function [62] is given by:

$$F_S(y) = \ln\left[\frac{1 + \exp\left(\dfrac{\psi_s - V - \alpha_b - \psi_t}{\phi_T}\right)}{1 + \exp\left(\dfrac{\psi_s - V_{GB} - \alpha_b - \psi_t}{\phi_T}\right)}\right] \tag{42}$$

where $q \cdot \alpha_b$ is the difference between the conduction band edge and the electron quasi-Fermi potential, and the variable ψ_t reflects the fact that there are few electrons having a kinetic energy higher than a few $k \cdot T$. Specifically, $\psi_t = 0$ for $V_{ox} \geq 0$ and $\psi_t = -V_{ox} + G_0 \cdot \phi_T$ for $V_{ox} < 0$, where G_0 is an adjustable parameter accounting for the possibility of a difference between the conduction band offset at the Si/SiO_2 and poly-Si/SiO_2 interfaces. In contrast to more empirical models, the use of the supply function F_S automatically ensures that gate current is zero for zero applied bias.

In the following we briefly discuss the gate-to-channel and the gate-overlap current components separately.

[6] In p-type MOSFETs, on the other hand, mainly valence band tunneling is important. In the following, the same derivation can be used for p-type MOSFETs but a different value for oxide potential barrier χ_B has to be used, see Figure 9.

Gate-to-Channel Current: The total contribution I_{GC} of the channel region to the gate-tunneling current is given by:

$$I_{GC} = W \cdot \int_0^L J_G(y) \cdot \mathrm{d}y \tag{43}$$

In order to calculate the above integral, the current continuity equation has to be solved:

$$\frac{\partial I_{DS}(y)}{\partial y} = -W \cdot J_{GC}(y) \tag{44}$$

where I_{DS} is given by (11) and is no longer constant along the channel. Eq. (44) cannot be solved explicitly, and as a consequence it needs to be approximated for compact modeling purposes. The current continuity equation is solved under the assumption that J_{GC} only induces a small perturbation of the potential distribution along the channel (i.e., $\partial I_{\mathrm{DS}}/\partial x \approx 0$). We note in passing that this assumption implies that I_{DS} is (approximately) constant along the channel and all equations derived in Section 2 are still valid. For this case, using the symmetric linearization method described in Section 2.4, Eq. (43) results in:

$$I_{GC} = I_{\mathrm{GINV}} \cdot F_S(y_m) \cdot D(y_m) \cdot p_{gc} \tag{45}$$

where I_{GINV} is theoretically given by $J_0 \cdot W \cdot L$ but is considered as an empirical parameter, y_m is the lateral coordinate of the surface potential midpoint as given by (29), and p_{gc} is a function of ψ_m and $\Delta\psi$. The latter can be found in the PSP documentation [32].

The total gate-to-channel current I_{GC} partitions into a source (I_{GCS}) and a drain component (I_{GCD}). Following [21]

$$I_{\mathrm{GCD}} = \frac{W}{L} \cdot \int_0^L y \cdot J_G(y) \cdot \mathrm{d}y \tag{46}$$

and $I_{\mathrm{GCS}} = I_{GC} - I_{\mathrm{GCD}}$. Again using the symmetric linearization method, the above integral results in

$$I_{\mathrm{GCD}} = I_{\mathrm{GINV}} \cdot F_S(y_m) \cdot D(y_m) \cdot p_{gd} \tag{47}$$

where p_{gd} is a function of ψ_m and $\Delta\psi$, which can be found in the PSP documentation [32].

Gate-Overlap Current: Essentially the same model for the tunneling current is used in both the channel and the overlap regions. However, in the latter case the position dependence of the surface potential is negligible, and hence the tunneling current density is approximately uniform. As a consequence, the gate-overlap current I_{Gov} in an overlap region with applied gate bias V_{GX} and surface potential ψ_{ov} is written as:

$$I_{\mathrm{Gov}} = I_{\mathrm{GOV}} \cdot F_S(\psi_{ov}, V_{GX}) \cdot D(z_{\mathrm{gov}}) \tag{48}$$

where I_{GOV} is theoretically equal to $J_0 \cdot W \cdot L_{ov}$, L_{ov} is the length of the gate/source or gate/drain overlap region, and z_{gov} is equal to V_{ov}/χ_B. The above equation is used for both gate-source and gate-drain overlap current by making V_{GX} equal to V_{GS} or V_{GD}, respectively.

Including the above components, the model gives an accurate description of gate current over the whole operation region for both n- and p-channel devices, see Figure 10. The gate current model provides Gummel symmetry as well.

3.4. Extrinsic Charges

For short-channel transistors, a major part of the total input capacitance C_{GG} is determined by the gate-to-source and gate-to-drain overlap capacitances. An accurate modeling of these bias-dependent overlap capacitances is thus important. Using Gauss' law and (34), the total charge in the overlap region is simply given by:

$$Q_{\mathrm{xov}} = \mathrm{CGOV} \cdot V_{ov} \tag{49}$$

where CGOV is a model parameter accounting for the geometry of the overlap region. Here again, X denotes either source or drain (with corresponding changes in ψ_{ov}). Taken together with the analytical approximation of ψ_{ov}

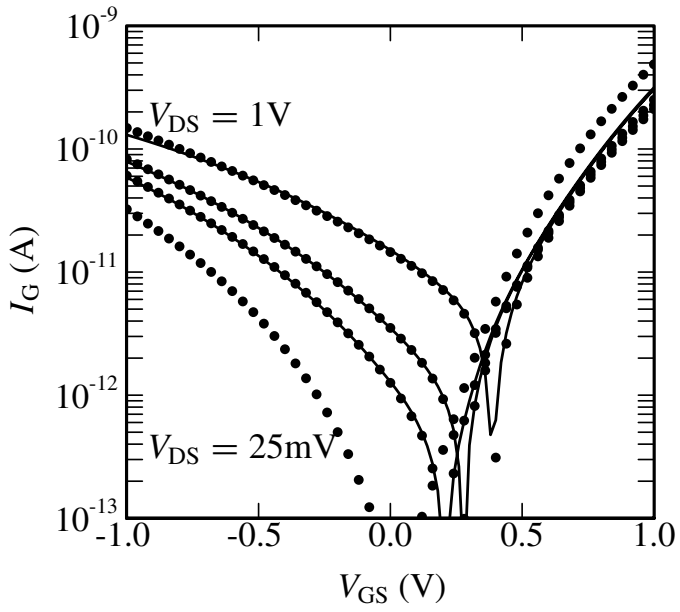


Figure 10. Gate current I_G versus gate-source bias V_{GS} at $V_{SB} = 0$ V and different drain-source bias V_{DS} for a $W/L = 360$ nm/90 nm n-channel MOSFET. Symbols denote measurements and lines represent modeled results using PSP.

illustrated in Figure 8, this expression provides a physical and computationally efficient description of the bias-dependent overlap charges eliminating the need for the mostly empirical modeling of Q_{xov} in older compact models.

In addition to the bias dependence of the overlap capacitance, the PSP model includes both the outer and inner-fringing charges (capacitances). The bias-independent outer fringing capacitance is a model parameter CFR and the outer fringing charge is simply CFR $\cdot$ V_{GX}. As described in [53] the inner fringing phenomena is strongly affected by the formation of the inversion layer and is consequently bias-dependent. In PSP inner fringing is modeled as the reduction of the source and drain terminal charges by ΔQ_S and ΔQ_D and corresponding change in the gate charge $\Delta Q_G = -\Delta Q_S - \Delta Q_D$ required to maintain the charge neutrality. Physically this reduction represents the deviation from the gradual-channel approximation inevitable in strong lateral-field regions close to the source and drain. Availability of ψ_{ov} enables formulation of the physically motivated semi-empirical expressions for ΔQ_S and ΔQ_D sufficient in engineering applications. Typical results for the extrinsic capacitances are shown in Figures 11. Further details including comparison with experimental data and two-dimensional simulations can be found in [11].

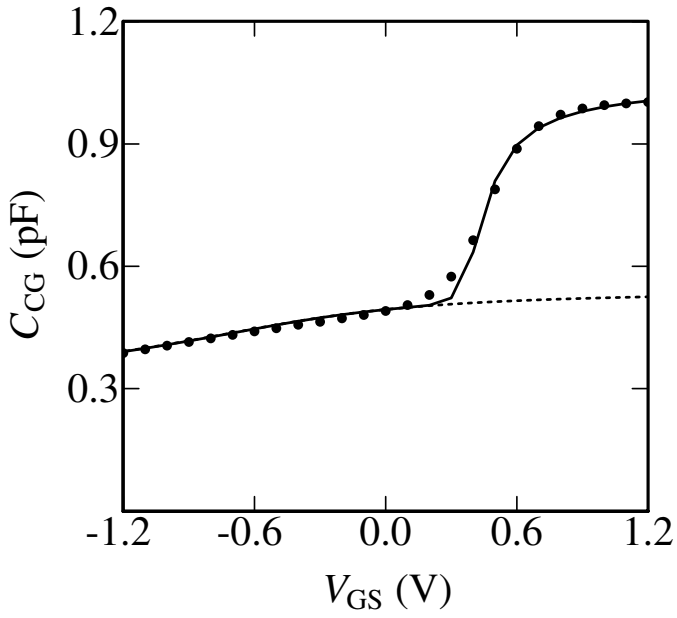


Figure 11. Channel-to-gate capacitance C_{CG} $(= C_{SG} + C_{DG})$ versus gate-source bias V_{GS} for short-channel n-type MOSFET; $V_{SB} = V_{DS} = 0\,\mathrm{V}$, $W/L = 800\,\mu\mathrm{m}/90\,\mathrm{nm}$. Symbols denote measurements, solid line denotes modeled extrinsic and intrinsic capacitances using PSP and dashed line denotes modeled extrinsic capacitance using PSP.

4. Noise Model

The circuit performance in analog and RF circuits is often limited by noise, and accurate modeling of noise behavior in circuit simulation is thus essential. In a MOSFET, generally three different types of noise can be observed: $1/f$ or flicker noise, thermal noise and induced gate noise. These types of noise are all related to the channel current. In reality, the gate tunnel current and the bulk current will also exhibit noisy behavior due to shot noise [63]. This has been taken into account in PSP as well, but is not further elaborated in this chapter.

In Section 4.1, the $1/f$ or flicker noise, as implemented in PSP, is briefly discussed. Since thermal noise and induced gate noise in a MOSFET stem from the same physical origin, they will both be treated in Section 4.2.

4.1. Flicker or 1/*f* Noise

At low frequencies, flicker or $1/f$-noise becomes dominant in MOSFETs. In the past, this type of noise was interpreted either in terms of trapping and detrapping of charge carriers in the gate oxide or in terms of mobility fluctuations. A general $1/f$-noise model by Hung *et al.* which combines both number and mobility fluctuations [64, 65], has found wide acceptance in the field of MOS modeling. The model assumes that the carrier number in the channel fluctuates due to trapping/detrapping of carriers in the gate oxide, and that these number fluctuations also affect the carrier mobility resulting in (correlated) mobility fluctuations. The model was originally formulated for V_T-based models. The PSP flicker noise model is obtained by developing a surface-potential-based version of the general model in [64, 65] resulting in an accurate expression for all operating regions. This formulation further develops an earlier version of the surface-potential-based adaption of [64, 65] given in [11, 22].

4.2. Thermal Noise and Induced Gate Noise

Thermal (or Nyquist) noise is caused by the random thermal (or Brownian) motion of carriers. In a MOSFET, the random motion of carriers in the channel translate to a fluctuation in the channel current I_{DS} flowing between drain and source. The channel current thus exhibits a frequency-independent (or white) noise spectral density S_{id}. In addition, owing to capacitive coupling between gate and channel, the fluctuations in the channel also induce a noise current in the gate terminal at high frequencies. Hence, apart from the channel current thermal noise spectral density S_{id}, the high-frequency noise also consists of the induced gate noise spectral density S_{ig}, which increases with f^2. Since both

S_{id} and S_{ig} stem from the same noise origin, they are partly correlated with correlation coefficient c.

Most available noise models for MOSFETs such as, e.g., the well-known Van der Ziel model [66], make use of the so-called Klaassen-Prins approach [67]. This approach, however, does not accurately account for velocity saturation [68]. As a result these models are inaccurate for short-channel devices [24, 69], where in particular S_{ig} is underestimated. An improved Klaassen-Prins approach, which accurately accounts for velocity saturation, was developed in [24, 69] and is used in MM11, level 1102, and in PSP.

In this approach, the channel current spectral density can be written as:

$$S_{id} = N_d \cdot \int_{V_{SB}}^{V_{DB}} g_c^2(V) \cdot \mathrm{d}V \tag{50}$$

where $N_d = 4 \cdot k \cdot T \cdot I_{DS}^{-1} \cdot L_c^{-2}$, and g_c and L_c denote the corrected channel conductivity and channel length, respectively. For the velocity saturation expression (18) used in PSP:

$$g_c(V) = \frac{g_0^2(V)}{g(V)} \tag{51}$$

$$L_c = L \cdot \frac{\int_{V_{SB}}^{V_{DB}} g_c(V) \cdot \mathrm{d}V}{\int_{V_{SB}}^{V_{DB}} g(V) \cdot \mathrm{d}V} \tag{52}$$

Here $g_0(V)$ is the channel conductivity without velocity saturation:

$$g_0(V) = \mu_{\mathrm{eff}} \cdot W \cdot q_i(V) \tag{53}$$

and $g(V)$ is the channel conductivity (including velocity saturation):

$$g(V) = \frac{g_0(V)}{\sqrt{1 + \left(\mu_{\mathrm{eff}} \cdot E_y / v_{\mathrm{sat}}\right)^2}} \tag{54}$$

Note that the channel current I_{DS} is a simple function of channel conductivity: $I_{DS} = g(V) \cdot \mathrm{d}V/\mathrm{d}y$.

The gate current spectral density can be written as [24, 69]:

$$\begin{aligned} S_{ig} = N_g \cdot & \int_{V_{SB}}^{V_{DB}} g_c^2(V) \\ & \cdot \left(\int_{V_{SB}}^{V} g_c\left(V'\right) \cdot \left[q_g\left(V'\right) - q_g(V) \right] \cdot \mathrm{d}V' \right)^2 \cdot \mathrm{d}V \end{aligned} \tag{55}$$

where $N_g = N_d \cdot \omega^2 \cdot W^2 / I_{DS}^4$. The cross-correlation spectral density between gate and drain current is given by [24]:

$$\begin{aligned} S_{igid} = N_{gd} \cdot & \int_{V_{SB}}^{V_{DB}} g_c^2(V) \\ & \cdot \left(\int_{V_{SB}}^{V} g_c\left(V'\right) \cdot \left[q_g\left(V'\right) - q_g(V) \right] \cdot \mathrm{d}V' \right) \cdot \mathrm{d}V \end{aligned} \tag{56}$$

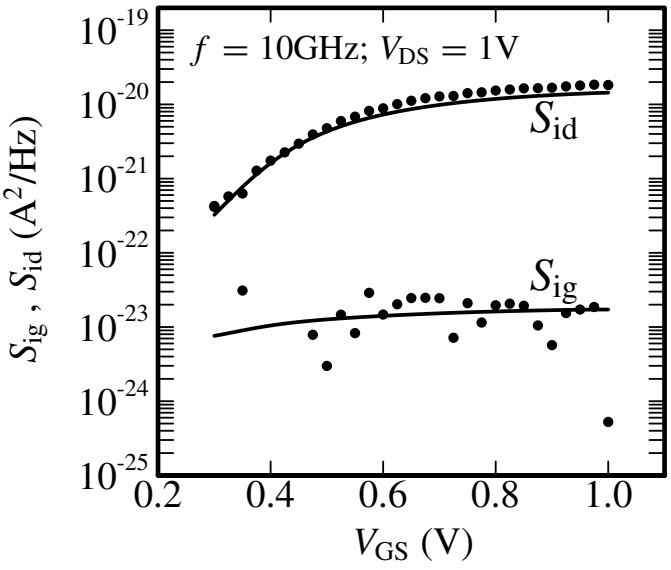


Figure 12. Drain (S_{id}) and gate (S_{ig}) current noise spectral density versus gate-source bias for an $L = 90\,\text{nm}$ n-channel device. Symbols denote measurements and lines represent modeled results using PSP.

where $N_{gd} = -j \cdot N_d \cdot \omega \cdot W/I_{DS}^2$. Finally, the correlation coefficient c is given by:

$$c = \frac{S_{igid}}{\sqrt{S_{ig} \cdot S_{id}}} \tag{57}$$

Using the symmetric linearization method, the improved Klaassen-Prins approach can be straightforwardly included in the ψ_s-framework. The corresponding expressions for S_{id}, S_{ig} and c can be found in the PSP documentation [32]. The resulting noise model gives an accurate description of high-frequency noise in MOSFETs down to deep-submicron dimensions, see Figure 12. The model is in good agreement with measurement data without using any additional noise parameters.

5. Junction Diode Model

In a MOS device, the drain/bulk and source/bulk junctions act as diodes, and as a result they will also contribute to the bulk current and capacitance. Due to the ever increasing junction steepness and pocket implantations, junction leakage is an increasing concern in CMOS technology scaling. The physical phenomena responsible for the increasing junction leakage are Shockley-Read-Hall

generation/recombination (SRH), trap-assisted tunneling (TAT) and band-to-band tunneling (BBT). Present-day compact models [1, 58] lack accurate physical descriptions of these effects. The PSP model contains a new junction diode model named JUNCAP2 [30], that is also available in stand-alone format. In contrast to earlier models [1, 58, 59], this model *(i)* gives single-piece expressions for SRH and TAT, valid in both forward and reverse mode of operation, *(ii)* removes the need for introducing an unphysical ideality factor, *(iii)* extends the existing model for TAT, valid at low fields, to the high-field regime encountered in modern MOS junctions, and *(iv)* is valid for junctions of arbitrary grading coefficient. In addition, the model incorporates shot noise in the junction current.

For the accurate modeling of a typical drain/bulk or source/bulk junction region, JUNCAP2 distinguishes three components: the bottom-edge, the STI-edge and the gate-edge component. These components scale differently with geometry, and, due to different junction steepness and doping concentrations at the different edges, these components show different electrical behavior. This is incorporated in JUNCAP2. As a result, JUNCAP2 gives an accurate description of the electrical behavior of junctions in modern CMOS technologies over a wide range of bias, geometry and temperature [30], see Figure 13.

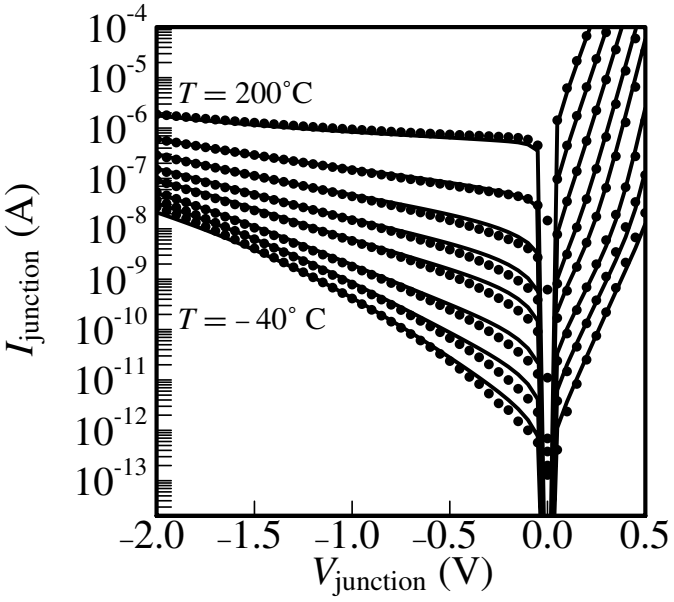


Figure 13. Junction leakage current I_{junction} versus applied junction bias V_{junction} at different temperatures for a typical n^+/p junction in 0.12 µm CMOS technology. Symbols denote measurements and lines represent modeled results using JUNCAP2.

6. Non-Quasi-Static Model

The intrinsic charge model described in Section 2.4 is quasi-static (QS). The QS approach assumes that a charge Q_X can be attributed to a terminal X and that Q_X changes instantaneously with a changing terminal voltage. In other words, it assumes that carriers travel at infinite velocity, which is not physical. A finite carrier velocity results, for example, in a phase shift (or delay) between the channel current and the gate voltage. This phase shift is not taken into account in the QS approach. This implies that for applications at high frequencies (approaching the cut-off frequency of the device) or for applications subject to fast transients, errors have to be expected in the QS approach due to non-quasi-static (NQS) effects. An NQS model of the MOSFET is thus essential for these applications.

Of the several NQS models developed at present, two allow an arbitrary trade-off between model accuracy and complexity: the channel segmentation method [70] and the spline-collocation technique [8, 15, 31]. The latter is more calculation-time efficient and is adopted in PSP after careful verification based in part on the channel segmentation method [31].

The spline collocation technique converts the partial differential equation expressing channel current continuity into a system of coupled ordinary differential equations that can be readily solved by circuit simulators. This is done as follows. Using (8) the continuity equation for the channel current $i(y,t)$

$$\frac{\partial i(y,t)}{\partial y} = W \cdot \frac{\partial q_i(y,t)}{\partial t} \tag{58}$$

is brought into a form [71] $R(y,t) = 0$ where:

$$R(y,t) = \frac{\partial q_i}{\partial t} + \frac{\partial}{\partial y} \cdot \left[\mu \cdot \left(\frac{q_i}{\mathrm{d}q_i/\mathrm{d}\psi_s}\right) - \phi_T\right] \cdot \frac{\partial q_i}{\partial y} \tag{59}$$

This automatically includes both drift and diffusion components of the current in the NQS model and with a proper choice of the $q_i(\psi_s)$ dependence includes all regions of MOSFET operation [15]. The collocation method is a particular form of the weighted residuals technique in which $q_i(y,t)$ dependence is approximated by a simpler function $q_a(y,t)$ and instead of demanding $R(y,t) = 0$ one imposes a weaker set of N conditions:

$$\int_0^L w_k(y) \cdot R_a(y,t) \cdot \mathrm{d}y = 0;\ k = 1, 2, \ldots, N \tag{60}$$

where $w_k(y)$ are appropriately chosen weighting factors and R_a is obtained from R by changing q_i into q_a. Specifically, for the collocation method

$$w_k = \delta(y - y_k) \tag{61}$$

where $y_k = k/(N+1)$. This is equivalent to requiring the continuity equation to be satisfied at N equidistant collocation points y_k rather than at any point along the channel.

A simple choice for q_a is a polynomial

$$q_a = \sum_{n=1}^{m} a_n(t) \cdot y^n \tag{62}$$

with time-dependent coefficients. This approach (with $N = 1$ and $m = 2$) has been used in the first successful application of the collocation method to the MOSFET NQS modeling [72]. Unfortunately, for $m > 2$ the polynomial approximation introduces unphysical oscillations of the inversion charge as a function of distance. This limits the technique to a single collocation point ($N = 1$) which is not sufficient, for example, for RF simulations and some fast transients.

A more powerful technique, the so-called spline collocation method, is to approximate the inversion charge by cubic splines with time-dependent coefficients selected as to provide continuity of q_a and its first two derivatives with respect to coordinate y. In this case q_a is oscillation-free for an arbitrary number of collocation points. Using Eqs. (60) and (61) one obtains a system of N ordinary first degree differential equations of the type

$$\frac{\mathrm{d}z_k}{\mathrm{d}t} = f_k(z_1, \ldots, z_k) \tag{63}$$

where $z_k = q_a(y_k, t)$ and f_k are known functions. Equations (63) are easily solved by circuit simulators (e.g., using coupled RC subcircuits) and the terminal currents are evaluated in terms of z_k and their time derivatives. Complete details are given in [8, 15, 31]. Here we note only that all terminal currents are automatically included in this approach. The NQS model used in PSP directly includes mobility reduction, velocity saturation and other small-geometry effects [31].

An important advantage of the spline collocation method is the arbitrary number of collocation points that translates into an arbitrary precision of the calculations (naturally, increasing N requires longer simulation times). Typically $N = 2$ is sufficient for transient simulations while $N = 5$ is used in RF applications. The latter also requires inclusion of the substrate subcircuit as described in [70]. Typical results for transient simulations are shown in Figure 14.

In addition to the overall reduction of the current, mobility reduction lengthens the transients. An example of RF simulations is shown in Figure 15 indicating a good agreement with measured results and channel segmentation method. In addition, PSP NQS model has been verified by comparison with the direct numerical solution of $R(y, t) = 0$. Since both the large-signal and small-signal NQS models use the same set of equations (63), they are consistent with each other and with quasi-static simulations, which appear as a proper limiting case

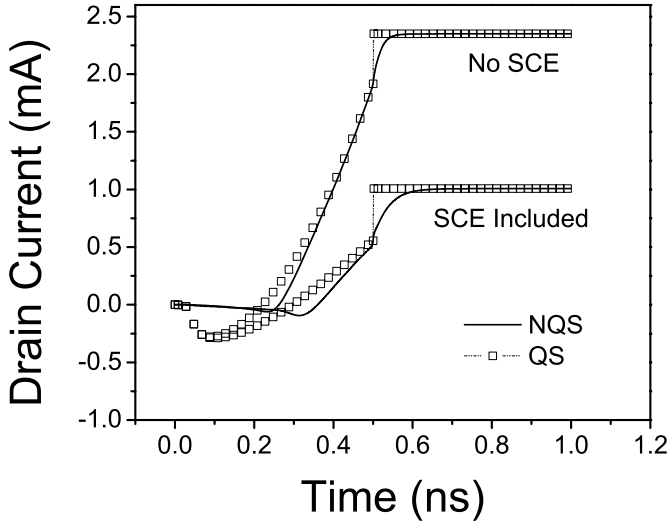


Figure 14. Transient response of $W/L = 5\,\mu\text{m}/5\,\mu\text{m}$ MOSFET with and without short-channel effects (SCE). The gate voltage is ramped from 0 to 3 V in 0.5 ns.

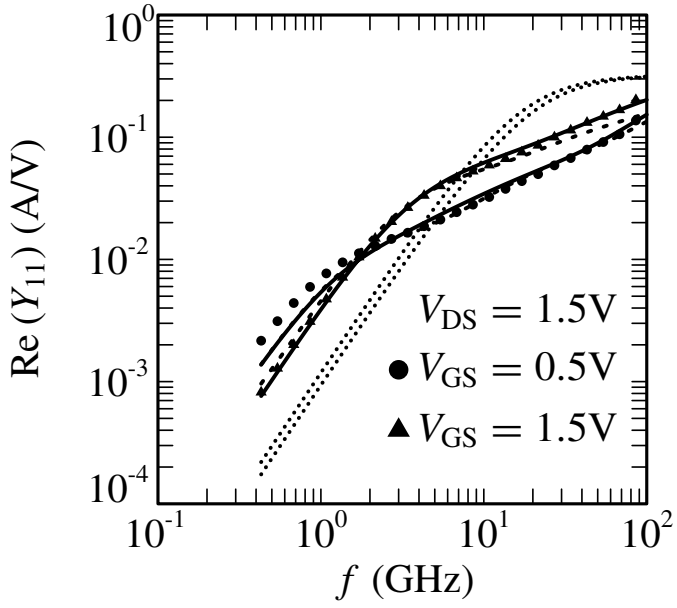


Figure 15. Real part of input admittance Y_{11} versus frequency f for different bias conditions for an n-channel MOSFET; $V_{SB} = 0$V, $W/L = 120\,\mu\text{m}/3\,\mu\text{m}$. Symbols denote measurements, dotted lines denote modeled results using PSP QS-model, solid lines denote modeled results using PSP NQS-model with $N = 5$, and dashed lines denote modeled results using $N = 5$ segmentation model [70] based on MM11. In the simulations bulk and gate resistances have also been taken into account.

of slow transients or in the low-frequency limit. This is not necessarily true for other NQS models.

7. Conclusions

The PSP model is a new compact MOSFET model which combines and extends the best features of the SP and MM11 models. The merger of SP and MM11 into PSP was facilitated by the compatibility of SP and MM11; both models are surface-potential-based, make use of some sort of symmetric linearization and make a distinction between local and global parameter level.

PSP is based on the formulation of surface potential and makes use of an analytical approximation of surface potential with an accuracy better than 1nV for both positive and negative bias on the source-bulk drain-bulk junctions. The derivation of the model expressions is considerably facilitated by the use of the symmetric linearization method. This method was developed in the framework of the SP model, and it has been expanded for PSP in order to include the velocity saturation model of MM11. It results in simple yet accurate expressions for the electrical quantities of the intrinsic MOS device, such as drain-source current, gate current, terminal charges and noise.

The extrinsic model in PSP includes accurate expressions for the gate current, the bulk current due to impact-ionization and gate-induced drain leakage, and the bias-dependent overlap capacitances. For this purpose, PSP uses a description of surface potential in the overlap regions, which is simpler than the above surface-potential description in the intrinsic region.

The noise model in PSP includes flicker noise, thermal noise, induced gate noise, and shot noise in the gate and bulk currents. The thermal noise and induced gate noise are partly correlated, and, in contrast to other models, their description accurately incorporates the impact of velocity saturation. The resulting noise model gives an accurate description of noise in MOS devices down to deep submicron devices without the use hot electron effects.

In addition, PSP contains a new junction diode model JUNCAP2, which is more accurate than state-of-the-art junction diode models. JUNCAP2 includes an accurate description of the Shockley-Read-Hall generation/recombination, trap-assisted tunneling and band-to-band tunneling phenomena, which are important in present-day and future CMOS technologies.

PSP incorporates a support module for the modeling of non-quasi-static (NQS) effects, which is important for high-frequency IC-design. The NQS-model in PSP makes use of the spline-collocation technique, which allows for a trade-off between complexity and model accuracy by changing the number of collocation points. In contrast to other NQS-models, this technique is suitable for both small-signal and large-signal simulations, and it is compatible with the quasi-static description in the limiting cases of slow transient and low-frequency

operation. The spline-collocation technique is less computation-time intensive than the channel segmentation method.

The PSP model has been subjected to the standard convergence tests and verified by comparison with data obtained from several 90 nm and 65 nm node processes. PSP has been selected as a new industry standard for the next generation compact MOSFET model by the Compact Modeling Council (CMC) [73].

Acknowledgments

The authors are particularly grateful to G.D.J. Smit, A.J. Scholten and D.B.M. Klaassen from Philips Research and H. Wang, W. Wu and X. Li from PSU. They would also like to thank the SiMKit-team from ED&T, L. Lemaître, C. McAndrew and G. Coram for their invaluable help with model implementation. Additionally, they would like to thank N. Arora, P. Bendix, D. Foty, W. Grabinski, A. Jha, C. McAndrew, S. Veeraraghavan, J. Victory, J. Watson and G. Workman for several illuminating discussions of the compact modeling methods, and to S. Hamm and B. Mulvaney for the use of the MICA simulator. Collaboration with T.-L. Chen on the development of the surface-potential-based models is gratefully acknowledged by one of the authors (GG).

PSP development at the PSU is supported in part by the Semiconductor Research Corporation, the Motorola Shared University Partnership Program and by the IBM University Partnership Award.

References

[1] BSIM3 and BSIM4: www-device.eecs.berkeley.edu

[2] Velghe, R.M.D.A.; Klaassen, D.B.M.; Klaassen, F.M. "MOS Model 9", *NL-UR 003/94*, Philips Electron. N.V., **1994**.
internet: www.semiconductors.philips.com/Philips_Models.

[3] Watts, J.; *et al.* "Advanced compact models for MOSFETs", In *Proc. NSTI-Nanotech*, **2005**, 3–12,

[4] Gildenblat, G.; Wang, H.; Chen, T.-L.; Gu, X.; Cai, X. "SP: An advanced surface-potential-based compact MOSFET model", *IEEE J. Solid-State Circ.*, **2004**, *39*, 1394–1406.

[5] Chen, T.-L.; Gildenblat, G. "Analytical approximation for the MOSFET surface potential", *Solid-State Electron.*, **2001**, *45*, 335–339.

[6] Chen, T.L.; Gildenblat, G. "Symmetric bulk charge linearisation in charge-sheet MOSFET model", *Electron. Lett.*, **2001**, *37*, 791–793.

[7] Gildenblat, G.; Chen, T.-L. "Overview of an advanced surface-potential-based model (SP)", In *Proc. NSTI-Nanotech*, **2002**, 657–661.

[8] Wang, H.; Chen, T.-L.; Gildenblat, G. "Quasi-static and nonquasi-static compact MOSFET models based on symmetric linearization of the bulk and inversion charges", *IEEE Trans. Electron Dev.*, **2003**, *50*, 2262–2272.

[9] Gu, X.; Wang, H.; Chen, T.L.; Gildenblat, G. "Substrate current in surface-potential-based models", In *Proc. NSTI-Nanotech*, **2003**, 310–312.

[10] Gildenblat, G.; Chen, T.-L.; Gu, X.; Wang, H.; Cai, X. "SP: An advanced surface-potential-based compact MOSFET model", In *Proc. CICC*, **2003**, 233–240.

[11] Gildenblat, G.; Cai, X.; Chen, T.-L.; Gu, X.; Wang, H. "Reemergence of the surface-potential-based compact MOSFET models", In *IEDM Tech. Digest*, **2003**, 863–866

[12] Gu, X.; Chen, T.-L.; Gildenblat, G.; Workman, G.O.; Veeraraghavan, S.; Shapira, S.; Stiles, K. "A surface potential-based compact model of n-MOSFET gate-tunneling current", *IEEE Trans. Electron Dev.*, **2004**, *51*, 127–135.

[13] Gildenblat, G.; McAndrew, C.C.; Wang, H.; Wu, W.; Foty, D.; Lemaitre, L.; Bendix, P. "Advanced compact models: Gateway to modern CMOS design", In *Proc. ICECS*, **2004**, 638–641.

[14] Wu, W.; Chen, T.-L.; Gildenblat, G.; McAndrew, C.C. "Physics-based mathematical conditioning of the MOSFET surface potential equation", *IEEE Transactions on Electron Dev.*, **July 2004**, *51*, 1196–1200.

[15] Wang, H.; Gildenblat, G. "A robust large signal non-quasi-static MOSFET model for circuit simulation", In *Proc. IEEE CICC*, **2004**, 5–8.

[16] Chen, T.-L.; Gildenblat, G. "An extended analytical approximation for the MOSFET surface potential", *Solid-State Electron.*, **2005**, *49*, 267–270.

[17] Wu, W.; *et al.*, "SP-SOI: A third generation surface potential based compact SOI MOSFET model", In *Proc. IEEE CICC*, **2005**, 819–822.

[18] van Langevelde, R.; Klaassen, F.M. "An explicit surface-potential-based MOSFET model for circuit simulation", *Solid-State Electron.*, **2000**, *44*, 409–418.

[19] van Langevelde, R.; Tiemeijer, L.F.; Havens, R.J.; Knitel, M.J.; Roes, R.F.M.; Woerlee, P.H.; Klaassen, D.B.M. "RF-distortion in deep-submicron CMOS technologies", In *IEDM Tech. Digest*, **2000**, 807–810.

[20] van Langevelde, R.; Scholten, A.J.; Havens, R.J.; Tiemeijer, L.F.; Klaassen, D.B.M. "Advanced compact MOS modeling", In *Proc. ESSDERC*, **2001**, 81–88.

[21] van Langevelde, R.; Scholten, A.J.; Duffy, R.; Cubaynes, F.N.; Knitel, M.J.; Klaassen, D.B.M. "Gate current: Modeling, ΔL extraction and impact on RF performance", In *IEDM Tech. Digest*, **2001**, 289–292.

[22] van Langevelde, R.; Scholten, A.J.; Klaassen, D.B.M. "MOS Model 11, level 1101", *NL-UR 2002/802*, Philips Electron. N.V., **2002**. www.semiconductors.philips.com/Philips_Models/mos_models/model11/

[23] van Langevelde, R.; Scholten, A.J.; Klaassen, D.B.M. "Physical background of MOS Model 11, level 1101", *NL-UR 2003/00238*, Philips Electron. N.V., **2003**. www.semiconductors.philips.com/Philips_Models/mos_models/model11/

[24] van Langevelde, R.; Paasschens, J.C.J.; Scholten, A.J.; Havens, R.J.; Tiemeijer, L.F.; Klaassen, D.B.M. "New compact model for induced gate current noise", In *IEDM Tech. Digest*, **2003**, 867–870.

[25] van Langevelde, R.; Scholten, A.J.; Klaassen, D.B.M. "Recent enhancements of MOS model 11", In *Proc. NSTI-Nanotech*, **2004**, 60–65.

[26] Klaassen, D.B.M.; van Langevelde, R.; Scholten, A.J. "Compact CMOS modeling for advanced analog and RF applications", *IEICE Trans. Electron.*, **2004**, *E87–C*, 854–866.

[27] Pao, H.C.; Sah, C.T. "Effects of diffusion current on characteristics of metal-oxide (Insulator)-semiconductor transistors", *Solid-State Electron.*, **1966**, *9*, 927–937.

[28] Brews, J.R. "A charge-sheet model of the MOSFET", *Solid-State Electron.*, **1978**, *21*, 345–355.

[29] Tsividis, Y.P. *Operation and modeling of the MOS transistor*, New York: McGraw-Hill, **1999**.

[30] Scholten, A.J.; Smit, G.D.J.; Durand, M.; van Langevelde, R.; Dachs, C.J.J.; Klaassen, D.B.M. "A new compact model for junctions in advanced CMOS technologies", In *IEDM Tech. Digest*, **2005**, 209–212.

[31] Wang, H. *et al.* "Unified non-quasi-static MOSFET model for large-signal and small-signal simulations", In *Proc. IEEE CICC*, **2005**, 823–826.

[32] PSP: pspmodel.ee.psu.edu

[33] Arora, N.D. *MOSFET models for VLSI circuit simulation*, Wien: Springer-Verlag, **1993**.

[34] McAndrew, C.C.; Victory, J.J. "Accuracy of approximations in MOSFET charge models", *IEEE Trans. Electron Dev.*, **2002**, *49*, 72–81.

[35] Sah, C.T. "A history of MOS transistor compact modeling", In *Proc. NSTI-Nanotech*, **2005**, 437–390.

[36] Boothroyd, A.R.; Tarasewicz, S.W.; Slaby, C. "MISNAN - A physically based continuous MOSFET model for CAD applications", *IEEE Trans. Comput.-Aided Design*, **1991**, *10*, 1512–1529.

[37] Rios, R.; Murdanai, S.; Shih W.-K.; Packan, P. "An efficient surface potential solution algorithm for compact MOSFET models", In *IEDM Tech. Digest*, **2004**, 755–758.

[38] Miura-Mattausch, M. *et al.* "HiSIM: A MOSFET model for circuit simulation connecting circuit performance with technology," In *IEDM Tech. Digest*, **2002**, 109–112.

[39] Turchetti, C.; Masetti, G. "A CAD-oriented analytical MOSFET model for high-accuracy applications", *IEEE Trans. Comput.-Aided Design*, **1984**, *3*, 117–122.

[40] Bagheri, M.; Tsividis, Y. "A small-signal DC-to-high-frequency non-quasistatic model for four-terminal MOSFETs valid in all regions of operation", *IEEE Trans. on Electron Dev.*, **1985**, *32*, 2383–91.

[41] Howes, R. *et al.* "A charge-conserving silicon-on-sapphire SPICE MOSFET model for analog design", *IEEE Int. Symp. Circ. Systems*, **1991**, *4*, 2160–2163.

[42] Nguyen, T.N.; Plummer, J.D. "Physical mechanisms responsible for short channel effects in MOS devices", In *IEDM Tech. Digest*, **1981**, 596–599.

[43] Skotnicki, T.; Merckel, G.; Pedron, T. "The voltage-doping transformation: A new approach to modeling of MOSFET short-channel effects", *IEEE Electron Dev. Lett.*, **1988**, *9*, 109–112.

[44] Miura-Mattausch, M. "Analytical MOSFET model for quarter micron technologies", *IEEE Trans. Comput.-Aided Design*, **1994**, *13*, 610–615.

[45] Joardar, K.; Gullapulli, K.K.; McAndrew, C.C.; Burnham M.E.; Wild, A."An improved MOSFET model for circuit simulation", *IEEE Trans. Electron Dev.*, **1998**, *45*, 134–148.

[46] Van de Wiele, F. "A long-channel MOSFET model", *Solid-State Electron.*, **1979**, *22*, 991–987.

[47] Huang, C.-L.; Arora, N. "Characterization and modeling of the n- and p-Channel MOSFETs inversion-layer mobility in the range 25–125°C", *Solid-State Electron.*, **1994**, *37*, 97–103.

[48] Bendix, P.; Rakers, P.; Wagh, P.; Lemaitre, L.; Grabinski, W.; McAndrew, C.C.; Gu, X.; Gildenblat, G. "RF distortion analysis with compact MOSFET models", In *Proc. IEEE CICC*, **2004**, 9–12.

[49] Scharfetter; D.L.; Gummel, H.K. "Large-signal analysis of a silicon read diode oscillator", *IEEE Trans. Electron Dev.*, **1969**, *16*, 64–77.

[50] El-Mansy, Y.A.; Boothroyd, A.R. "A simple two-dimensional model for IGFET operation in the saturation region", *IEEE Trans. Electron Dev.*, **1977**, *24*, 254–262.

[51] Cao, K.M. *et al.* "Modeling of pocket implanted MOSFETs for anomalous analog behavior", In *IEDM Tech. Digest*, **1999**, 171–174.

[52] Ward, D.E.; Dutton, R.W. "A charge-oriented model for MOS transistor capacitances", *IEEE J. Solid-State Circ.*, **1978**, *13*, 703–708.
[53] Foty, D. *MOSFET Modeling with SPICE: Principles and Practice*, Upper Saddle River, NJ: Prentice-hall, **1997**.
[54] Liu, W. *MOSFET Models for SPICE Simulations Including BSIM3v3, BSIM4*, New York: Wiley, **2001**.
[55] Victory, J.; Yan, Z.; Gildenblat, G.; McAndrew, C.; Zheng, J. "A physically based scalable varactor model and extractor methodology for RF applications", *IEEE Trans. Electron Dev.*, **2005**, *52*, 1343–1353.
[56] Chen, J.; Chan, T.Y.; Ko, P.K.; Hu, C. "Subbreakdown drain leakage current in MOSFET", *IEEE Electron Dev. Lett.*, **1987**, *8*, 515–517.
[57] Kane, E.O. "Zener tunneling in semiconductors", *J. Phys. Chem. Solids*, **1959**, *12*, 181–188.
[58] JUNCAP level 1: www.semiconductors.philips.com/Philips_Models
[59] Hurkx, G.A.M.; de Graaff, H.C.; Kloosterman, W.J.; Knuvers, M.P.G. "A new analytical diode model including tunneling and avalanche breakdown", *IEEE Trans. Electron Dev.*, **1992**, *39*, 2090–2098.
[60] Wright, P.J.; Saraswat, K.C. "Thickness limitations of SiO_2 gate dielectrics for MOS ULSI", *IEEE Trans. Electron Dev.*, **1990**, *37*, 1884–1892.
[61] Choi, C.-H.; Nam, K.-Y.; Yu, Z.; Dutton, R.W. "Impact of gate direct tunneling current on circuit performance: A simulation study", *IEEE Trans. Electron Dev.*, **2001**, *48*, 2823–2829.
[62] Tsu, R.; Esaki, L. "Tunneling in a finite superlattice", *Appl. Phys. Lett.*, **1973**, *22*, 562–564.
[63] Scholten, A.J.; Tiemeijer, L.F.; van Langevelde, R.; Havens, R.J.; Zegers-van Duijnhoven, A.T.A.; Venezia, V.C. "Noise modeling for RF CMOS circuit simulation", *IEEE Trans. Electron Dev.*, **2003**, *50*, 618–632.
[64] Hung, K.K.; Ko, P.K.; Hu, C.; Cheng, Y.C. "A unified model for the flicker noise in metal-oxide-semiconductor field-effect transistors", *IEEE Trans. Electron Dev.*, **1990**, *37*, 654–665.
[65] Hung, K.K.; Ko, P.K.; Hu, C.; Cheng, Y.C. "A physics-based MOSFET noise model for circuit simulators", *IEEE Trans. Electron Dev.*, **1990**, *37*, 1323–1333.
[66] van der Ziel, A. *Noise Solid-State Dev. Circuits*, New York: Wiley-Interscience, **1986**.
[67] Klaassen, F.M.; Prins, J. "Thermal noise of MOS transistors", *Philips Res. Reports*, **1967**, *22*, 505–514.
[68] Klaassen, F.M. "Comments on hot carrier noise in field-effect transistors", *IEEE Trans. Electron Dev.*, **1971**, *18*, 74–75.
[69] Paasschens, J.C.J.; Scholten, A.J.; van Langevelde, R. "Generalisations of the Klaassen-Prins equation for calculating the noise of semiconductor Devices", *IEEE Trans. Electron Dev.*, **2005**, *52*, 2463–2472.
[70] Scholten, A.J.; Tiemeijer, L.F.; de Vreede, P.W.H.; Klaassen, D.B.M. "A large signal non-quasi-static MOS model for RF circuit simulation", In *IEDM Tech. Digest*, **1999**, 163–166.
[71] Mancini, P.; Turchetti, C.; Masetti, G. "A non-quasi-static analysis of the transient behavior of the long-channel MOSFET valid in all regions of operation", *IEEE Trans. Electron Dev.*, **1987**, *ED-34*, 325–334.
[72] Hwang, S.W.; Yoon, T.-W.; Kwon, D.H.; Kim, K.H. "A physics-based SPICE-compatible non-quasi-static MOS transient model for RF circuit simulation", *Jpn. J. Appl. Phys.*, **1998**, *37*, L119–L121.
[73] CMC-website: www.eigroup.org/cmc

Chapter 3

EKV3.0: AN ADVANCED CHARGE BASED MOS TRANSISTOR MODEL

A Design-oriented MOS Transistor Compact Model for Next Generation CMOS

Matthias Bucher, Antonios Bazigos*, François Krummenacher**, Jean-Michel Sallese**, and Christian Enz**

Technical University of Crete (TUC), 73100 Chania, Crete, Greece
E-mail: bucher@electronics.tuc.gr
**National Technical University of Athens (NTUA), 15773 Athens, Greece*
***Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland*

Abstract: The EKV3.0 MOS transistor compact model addresses the design and circuit simulation of analog, digital and RF integrated circuits using advanced sub-100 nm CMOS technologies. This chapter presents the physical foundation of the charge model, as well as its extensions to account for geometrical effects, gate current, noise etc. The model is compared to data ranging from 0.25 um to 90 nm CMOS generations. A parameter extraction procedure is outlined. EKV3.0 has been developed in the Verilog-A behavioral language for reasons of portability among simulators.

Key words: MOS transistor; compact model; next generation; nanoscale CMOS; weak inversion; moderate inversion; analog/RF circuit design; EKV model; Verilog-A.

1. Introduction

For circuit-level design of CMOS analog and radio frequency integrated circuits (RFICs), the compact MOS transistor (MOST) model is the key "workhorse" enabling the designer to efficiently achieve design goals. Recently, the demand from the circuit design community for highly consistent,

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, 67–95.

physics-based and full-featured compact models has increased particularly in view of using sub-100 nm CMOS technologies.

A primary concern for advanced MOST models is its physical basis. The charge-based model approach taken within the EKV model is itself based on a surface-potential analysis. The basic charge modelling approach [3–12] allows physically consistent and accurate modelling of current, terminal charges and noise, without introducing artificial parameters besides the physical parameters of surface potential modeling (e.g. [13, 14]). Besides supporting full circuit simulation, the compact model should however also have an efficient counterpart for circuit design. The development of the EKV model always was driven by the needs of analog IC design [1, 2]. For many circuit applications, even at RF frequencies, operation in weak and particularly moderate inversion may offer a favorable trade-off among power consumption, linearity, matching, noise and bandwidth. The charge-based approach offers suitable expressions for hand-calculation, which a surface-potential only model cannot offer.

For advanced CMOS generations, new effects have appeared which have a significant impact on circuit design, such as undesirable gate tunneling currents, layout-dependent stress effects affecting each device as well as geometrical scaling and many more. Analog circuit design requires particular attention for accurate modelling of transconductances over all bias ranges and geometries [36, 37]. For applications at radio-frequencies (RF), multi-finger device layout is commonly used [27, 30, 33, 35], which combines the above-mentioned effects with the complexity of non-quasistatic (NQS) behavior of the MOS channel [28, 29, 31]: the traditional quasistatic approach for handling the MOS channel is insufficient to accurately account for high-frequency effects. Thermal noise in short-channel transistors is enhanced [34], while at high frequencies, channel thermal noise is capacitively coupled into gate and substrate (induced gate and substrate noise) [32].

The present chapter presents the basic approach taken in the context of the EKV MOST model to implement the above effects. The full-featured EKV3.0 compact MOST model [41–47] for circuit simulation is presented together with its basic list of parameters. Application examples range from 0.25 um to 90 nm CMOS. A parameter extraction procedure is outlined [38–41], and implementation in Verilog-A language [48, 49] is shortly discussed.

2. Ideal Charge-Based Model of the MOS Transistor

2.1. Surface Potential and Inversion Charge Modelling

The total channel charge density Q'_C in an infinitesimal piece of the channel is found by applying Gauss' law,

$$Q'_C = -C'_{OX} \cdot (V_G - V_{FB} - \Psi_S) \tag{1}$$

where Ψ_S is the surface potential, $C'_{OX} = \varepsilon_{OX}/T_{OX}$ the oxide capacitance per unit area, and V_{FB} the flat-band voltage. The bulk depletion charge Q'_B is given by,

$$Q'_B = -\sqrt{2q\varepsilon_{si}N_{\text{sub}}\Psi_S} \tag{2}$$

and ε_{OX} and ε_{si} are the permittivities of silicon and silicon dioxide, respectively. The gate oxide thickness T_{OX} and the substrate doping concentration N_{sub}, together with V_{FB} are the main actual physical parameters describing the MOS technology.

Inversion charge is then expressed as,

$$Q'_I = Q'_C - Q'_B = -C'_{OX} \cdot \left(V_G - V_{FB} - \Psi_S - \gamma\sqrt{\Psi_S}\right) \tag{3}$$

where $\gamma = \sqrt{2q\varepsilon_{si}N_{\text{sub}}}/C'_{OX}$ is the substrate effect parameter. As can be seen in Figure 1, the relation among inversion charge and surface potential at fixed gate voltage is approximately linear. Linearizing the inversion charge versus surface potential provides the inversion charge linearization factor n_q,

$$n_q \equiv \frac{\partial\left(Q'_I/C'_{OX}\right)}{\partial\Psi_S} = 1 + \frac{\gamma}{2\sqrt{\Psi_S}} \tag{4}$$

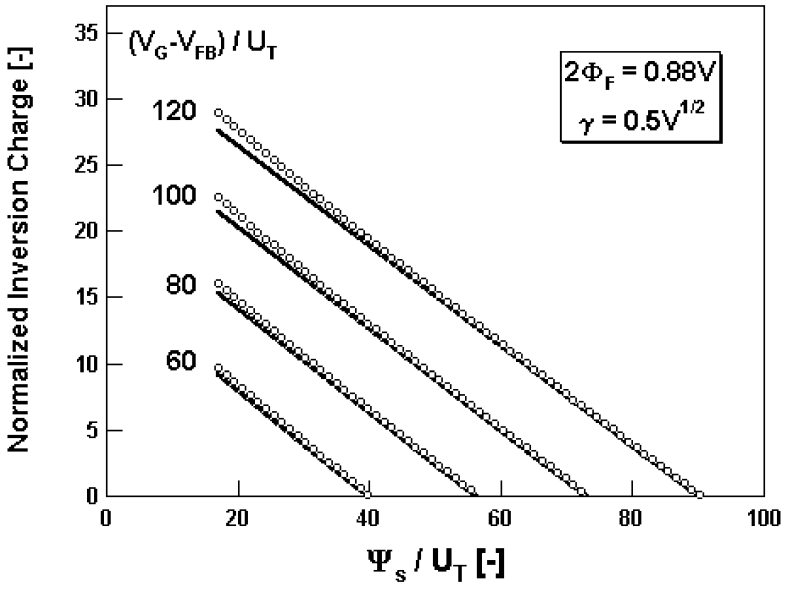


Figure 1. Normalized inversion charge versus surface potential Ψ_S for varied, fixed values of gate voltage V_G. Numerically calculated (markers) and approximation by linearization (lines).

Neglecting, on the other hand, inversion charge density in (3) provides the pinch-off surface potential Ψ_P [5, 10, 12, 15],

$$\Psi_P \equiv \Psi_S|_{Q_I=0} = V_G - V_{FB} + \gamma \cdot \left[\frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} + V_G - V_{FB}}\right] \tag{5}$$

We can therefore express the inversion charge as,

$$Q'_I \cong n_q \cdot C'_{OX} \cdot (\Psi_S - \Psi_P) \tag{6}$$

We then define the pinch-off voltage V_P as [12],

$$V_P \equiv \Psi_P - \Psi_0 \quad \text{where} \quad \Psi_0 \cong 2\Phi_F = 2U_T \ln\left(\frac{n_i}{N_{\text{sub}}}\right) \tag{7}$$

where Φ_F is the quasi-Fermi potential and n_i the intrinsic carrier concentration.

A convenient approximation of the pinch-off voltage is [5],

$$V_P \cong \frac{V_G - V_{TO}}{n} \quad \text{where} \quad V_{TO} = V_{FB} + \Psi_0 + \gamma\sqrt{\Psi_0} \tag{8}$$

where n is the slope factor,

$$n \equiv \left[\frac{\partial \Psi_P}{\partial V_G}\right]^{-1} = 1 + \frac{\gamma}{2\sqrt{\Psi_P}} \tag{9}$$

An illustration of pinch-off voltage and slope factor is given in Figure 2.

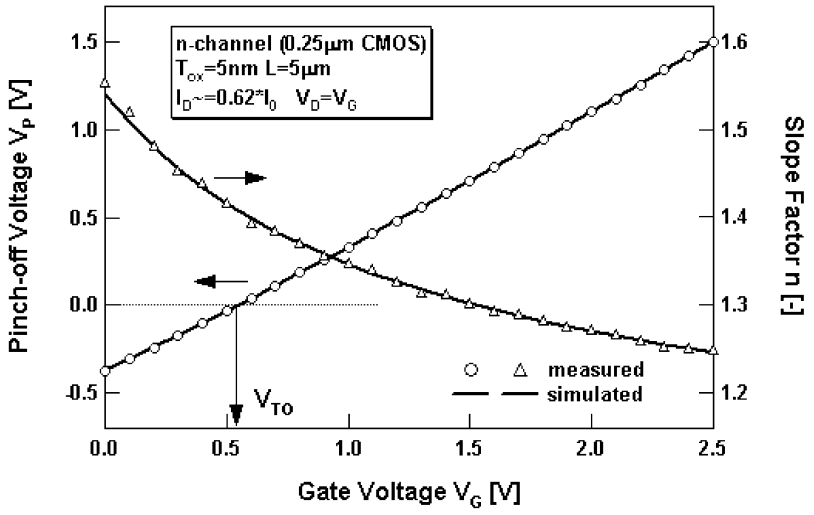


Figure 2. Pinch-off voltage (left axis) and slope factor (right axis) versus gate voltage, measurement and EKV3.0 model.

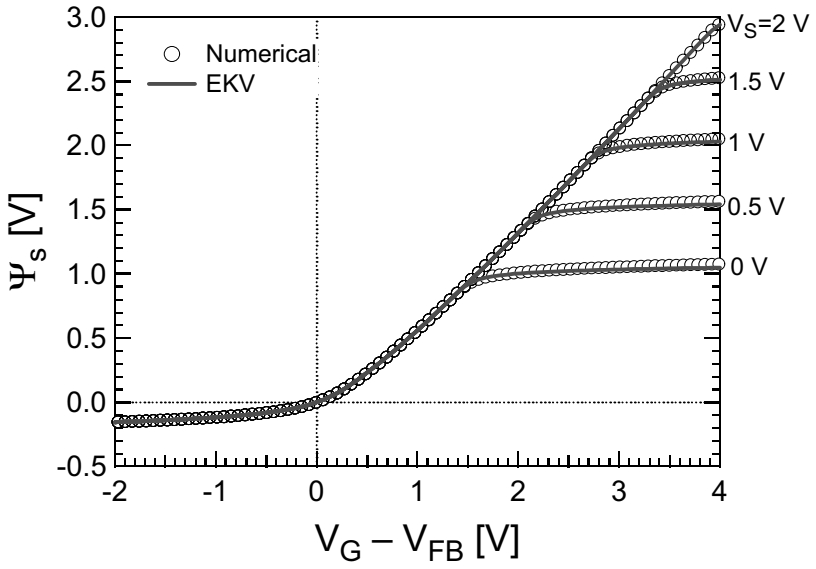


Figure 3. Surface potential Ψ_S versus gate voltage V_G and various values of V_S. The EKV model provides an accurate and continuous approximation to numerically calculated surface potential from accumulation through depletion and inversion.

Note that formally the inversion slope factor n_q and the slope factor n are very close. More detail on the interpretation of both can be found in []. While n_q appears in the normalization quantities for charges and current, the slope factor n is related to the substrate effect and hence they have a different role which needs to be kept separate in the model code for computer simulation. For approximate use in terms of hand calculation, both may be assumed the same.

The surface potential is not used explicitly in the model, but can be recalculated from the charge expressions. Anticipating the further model expressions for accumulation-, depletion- and inversion charge (see next section), Figure 3 shows the result of EKV3.0 modelling of surface potential versus gate voltage and different channel voltages. The model compares well with the numerical solution for the surface potential, and provides continuity through all modes of operation.

2.2. Model for Drain Current

The current transport equation in MOS transistors is written,

$$I_D = \mu \cdot W \cdot \left(-Q'_I \cdot \frac{\partial \Psi_S}{\partial x} + U_T \cdot \frac{\partial Q'_I}{\partial x} \right) \tag{10}$$

where μ is the carrier mobility. Using the charge linearization scheme [3–8, 10],

$$\frac{\partial \Psi_s}{\partial x} \cong \frac{1}{n_q} \frac{\partial Q_i'}{\partial x} \tag{11}$$

allows us to integrate the channel current I_D from source to drain in terms of source and drain inversion charge densities q_s and q_d, respectively [10, 12],

$$I_D = 2 \cdot n_q \cdot U_T^2 \cdot \mu \cdot C_{OX}' \frac{W}{L} \left[q_s^2 + q_s - q_d^2 - q_d \right] \tag{12}$$

Note in the above that the drain current can now be written in symmetric forward and reverse normalized currents i_f and i_r, [5] respectively,

$$I_D = I_{\text{Spec}} \cdot \left[i_f - i_r \right] \begin{cases} i_f = q_s^2 + q_s \\ i_r = q_d^2 + q_d \end{cases} \tag{13}$$

where I_{Spec} is the specific current [5],

$$I_{\text{Spec}} = 2 \cdot n_q \cdot \beta \cdot U_T^2 \quad \text{where} \quad \beta = \mu \cdot C_{OX}' \frac{W}{L} \tag{14}$$

The only missing relationship is the one linking charge to applied voltages. It can be shown that the following relationship among pinch-off voltage, inversion charge density and channel voltage v_{ch} holds throughout the channel [8, 10, 12],

$$v_P - v_{ch} = 2q_i + \ln(q_i) \begin{cases} v_P - v_S = 2q_s + \ln(q_s) \\ v_P - v_D = 2q_d + \ln(q_d) \end{cases} \tag{15}$$

This relationship clarifies the linear relationship among charge and voltage corresponding to strong inversion ($v_P - v_{S,D} > 0$), while the logarithmic relationship results in weak inversion ($v_P - v_{S,D} < 0$). From these relationships, it is easy to derive tables of approximate relationships for drain current and transconductances holding in weak/strong inversion, as well as saturation/non-saturation according to the relations among v_D and v_S.

Note that the above relationship is not analytically invertible to express charge in terms of voltage. This inversion is achieved by an approximation yielding high accuracy and continuity.

2.3. Transconductances

The relationship among transconductance and inversion charge densities at source and drain is immediate [5, 10],

$$\begin{matrix} g_{ms} = Y_{\text{Spec}} \cdot q_s \\ g_{md} = Y_{\text{Spec}} \cdot q_d \end{matrix} \quad \text{where} \quad Y_{\text{Spec}} = 2 \cdot n_q \cdot \beta \cdot U_T \tag{16}$$

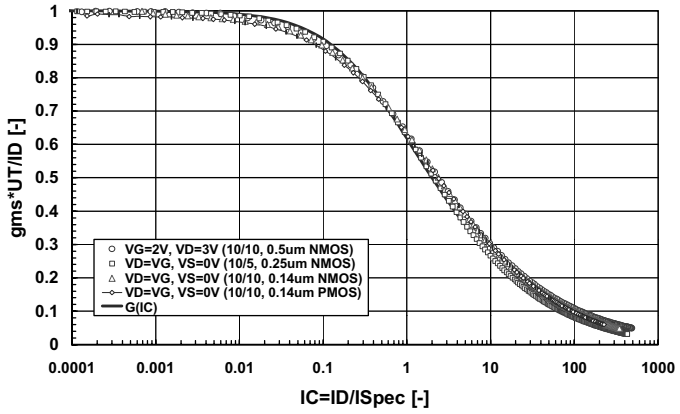


Figure 4. Normalized transconductance versus normalized current, from different NMOS and PMOS transistors from various CMOS technologies.

Noting the relationships among normalized current and charge, the important relationship among transconductance and normalized current is established [7],

$$\frac{g_{ms} \cdot U_T}{I_D} = \frac{1}{\frac{1}{2} + \sqrt{\frac{1}{4} + i_f}} \qquad \frac{g_{md} \cdot U_T}{I_D} = \frac{1}{\frac{1}{2} + \sqrt{\frac{1}{4} + i_r}} \tag{17}$$

Figure 4 shows normalized source transconductance of various transistors versus normalized current in saturation operation, compared to the above theoretical expression. Measurements coincide with the theory for a wide range of different CMOS technologies.

Further interesting relationships among different transconductances can be established [37],

$$g_m = \frac{g_{ms} - g_{md}}{n} \quad \text{and} \quad g_{mb} = \frac{n-1}{n}(g_{ms} - g_{md}) \tag{18}$$

2.4. Integral Charges and Transcapacitances

Integration of local charge densities along the MOS channel provides a means to express the total inversion and depletion charge. Ward's charge partitioning scheme [18] is applied to attribute a part of each channel charge to

either source or drain,

$$Q_I = W \cdot \int_0^L Q_I'(x) \cdot \mathrm{d}x \tag{19}$$

$$Q_D = W \cdot \int_0^L \frac{x}{L} Q_I'(x) \cdot \mathrm{d}x \tag{20}$$

$$Q_S = W \cdot \int_0^L \left(1 - \frac{x}{L}\right) Q_I'(x) \cdot \mathrm{d}x \tag{21}$$

where we note that $Q_I = Q_S + Q_D$. The transcapacitances are then obtained using partial differentiation,

$$C_{XY} \equiv \pm\delta \frac{\partial Q_X}{\partial V_Y} \quad \text{where} \quad \delta = \begin{cases} +1 & X = Y \\ -1 & else \end{cases} \tag{22}$$

An illustration of total gate capacitance, including polydepletion effect (see next section) is shown in Figure 5 versus gate and drain voltage. A notable difficulty is achieving continuous charge and transcapacitance expressions across the flat-band voltage. Further details can be found in [10, 19].

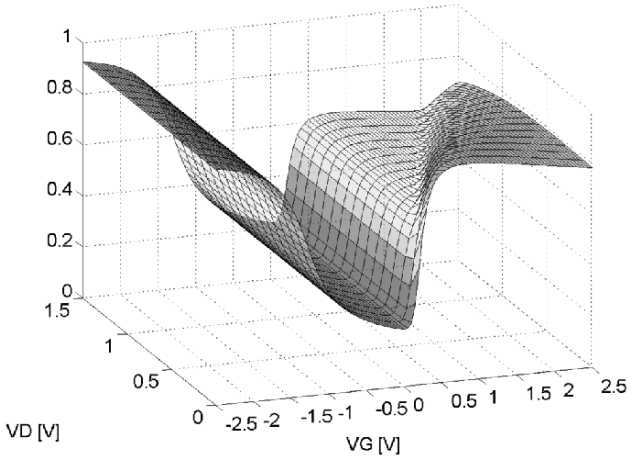


Figure 5. Normalized total gate transcapacitance C_{GG}.versus gate and drain voltages V_G.and V_D, where $V_S = 0$ V. The operating regions cover accumulation (left) to depletion and inversion (right), and linear operation (front) to saturation (back).

2.5. High-Frequency Model

A general model for high-frequency small-signal operation [28, 31]is shown in Figure 6. The three voltage controlled current sources (VCCS) are defined as,

$$\begin{aligned} I_m &= Y_m \cdot (V(gi) - V(bi)) \\ I_{ms} &= Y_{ms} \cdot (V(si) - V(bi)) \\ I_{md} &= Y_{md} \cdot (V(di) - V(bi)) \end{aligned} \tag{23}$$

General relationships hold in all operating regions among transadmittances and admittances,

$$\begin{aligned} Y_m &= \frac{Y_{ms} - Y_{md}}{n} \\ Y_{gbi} &= \frac{n-1}{n}(j\omega \cdot WLC'_{ox} - Y_{gsi} - Y_{gdi}) \\ Y_{bsi} &= (n-1) \cdot Y_{gsi} \\ Y_{bdi} &= (n-1) \cdot Y_{gdi} \end{aligned} \tag{24}$$

The above transadmittances are governed by a bias-dependent critical normalized frequency $\Omega_{\text{crit}} = \omega_{\text{crit}}/\omega_{\text{spec}}$ defined as [28, 31],

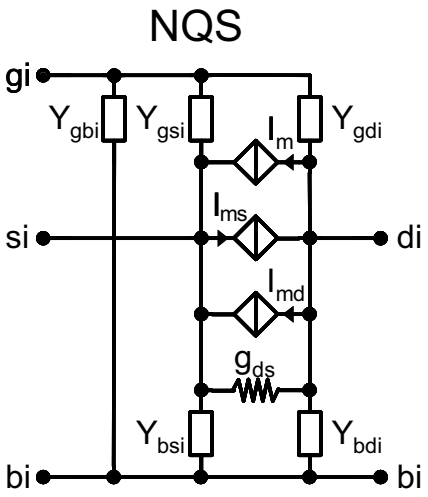


Figure 6. Small-signal equivalent circuit for HF application.

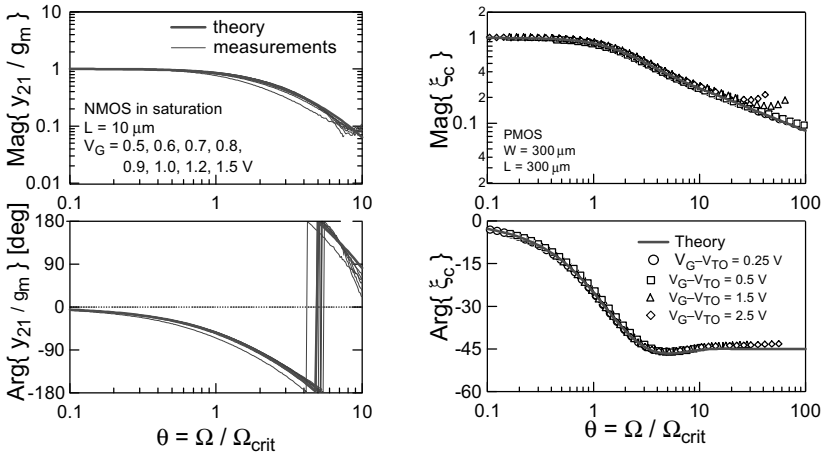


Figure 7. Magnitude and phase of normalized NQS small signal auxiliary functions versus normalized frequency, compared to measurement.

$$\Omega_{\text{crit}} = \frac{30 \cdot (q_s + q_d + 1)^3}{4q_s^2 + 4q_d^2 + 12q_s q_d + 10q_s + 10q_d + 5}$$
$$\Omega_{\text{crit}} = \begin{cases} \dfrac{15}{2} q_s & SI(sat.) \\ 6 & WI(sat.) \end{cases} \tag{25}$$

where $\omega_{\text{spec}} = \mu U_T / L^2$. The non-quasistatic model reduces to the quasistatic counterpart at lower frequencies, essentially depending on inversion conditions, besides mobility and channel length.

The 3 transadmittances and the 5 admittances depend on two general auxiliary functions, ξ_m and ξ_c, respectively. These are detailed in [28, 31] and further illustrated in Figure 7.

3. Extensions of Charge-Based Modelling Approach

The ideal MOS transistor model framework as presented in the previous section needs to be complemented to account for all imperfections related to high-field effects, high doping concentrations, thin gate dielectric, parasitic capacitances and leakage, series resistance etc. These effects are summarized in Table 1. Several among these will be further presented throughout the following subsections.

Table 1. Effects covered in the EKV3.0 compact MOS transistor model.

"Long-channel"	**"Short-/Narrow channel"**
Polydepletion (PD) effect	Reverse short-channel effect (RSCE)
Quantum mechanical (QM) effect	Inverse narrow width effect (INWE)
PD effect in accumulation in MOS varactors	Source/drain charge sharing
Continuous depletion/accumulation charge/transcapacitances	Drain induced barrier lowering (DIBL)
	Weak inversion slope degradation
Vertical/lateral non-uniform doping	Velocity saturation (variable order) channel length modulation
Vertical field dependent mobility based on effective field including Coulomb, phonon- and surface roughness scattering	Hot-carrier effects on short-channel thermal noise
Output conductance degradation due to pocket/halo implants	2nd order scaling effects
	Matching
NQS effects, consistent large- and small signal approach	**Parasitic effects**
Thermal noise, flicker noise	Bias-dependent series resistance
Induced gate- and substrate noise at NQS conditions.	Bias-dependent overlap & inner fringing charge/capacitance
	Gate tunnelling current
	Gate induced source/drain leakage
	Edge conduction effect

3.1. Polydepletion and Quantum Effects

Depletion in the polysilicon gate and energy quantization of the mobile carriers in the channel drastically reduce the performance of deep submicron CMOS technology. Quantum mechanical (QM) and polydepletion (PD) effects delay the formation of either accumulation or inversion charge with applied gate bias. The most immediately observed changes in device characteristics are increased threshold voltage and decreased gate capacitance, resulting in reduced drain current. Implementation of both these effects has been presented in [15–17].

Polydepletion, resulting from insufficient doping of the polysilicon gate, usually occurs when the MOS channel is inversion for usual type of gate doping, i.e. opposite to the type of channel doping. The EKV3.0 model provides however also the possibility of choosing the same doping type for the gate as for the channel. For further discussion of this point the reader is referred also to the section on overlap charge/capacitance as well as parameter extraction.

3.2. Mobility, Velocity Saturation and Channel Length Modulation

Various scattering mechanisms reduce carrier mobility depending on the field strength, either vertical field, or longitudinal field in the MOS channel. In long-channel MOSTs operating in inversion, the mobility of the carriers is dominated by Coulomb scattering at low vertical field, while phonon scattering dominates at intermediate and surface roughness scattering at high vertical field strength. A convenient way to combine these effects is via the Matthiessen rule,

$$\frac{1}{\mu} = \frac{1}{\mu_C} + \frac{1}{\mu_{ph}} + \frac{1}{\mu_{sr}} \tag{26}$$

where μ_C, μ_{ph} and μ_{sr} correspond to the three respective mobility effects.

Coulomb scattering increases at lower temperatures and higher doping densities, and is therefore important at low-temperature operation and/or with highly doped substrates as in advanced CMOS. Phonon-scattering has a well-known temperature dependence reducing mobility at higher temperatures also in less highly doped MOS channels and intermediate vertical field strength, while surface roughness scattering is only slightly temperature dependent.

Due to the field-dependence of the scattering mechanisms, mobility in the MOS channel is position dependent due to the change in charge density along the channel. An integration along the channel provides the integral mobility of the MOS transistor,

$$\bar{\mu} = \frac{1}{\frac{1}{L}\int_0^L \left[\frac{1}{\mu_C} + \frac{1}{\mu_{ph}} + \frac{1}{\mu_{sr}}\right] \cdot \mathrm{d}x} \tag{27}$$

where the Coulomb, phonon- and surface roughness scattering limited mobility terms depend on local charges or vertical field $E_\perp = \left|Q'_B + \eta Q'_I\right| / \varepsilon_{si}$, respectively, as,

$$\frac{1}{\mu_C} \propto \left[\frac{1}{2} + \frac{|Q'_I|}{\varepsilon_{si}}\right]^{(-1 \geq \alpha_C \geq -2)} \quad \frac{1}{\mu_{ph}} \propto [E_\perp]^{1/3} \quad \frac{1}{\mu_{sr}} \propto [E_\perp]^2 \tag{28}$$

The above integration can be carried out resulting in an expression notably depending on inversion charge densities at source and drain. The integral mobility is then used in the drain current expression. As a result, vertical field mobility is naturally dependent not only on gate, but also source and drain voltages, without introducing artificial dependences or parameters. An illustration of the resulting mobility for a long-channel transistor is shown in Figure 8.

In short-channel transistors, velocity saturation is the main effect limiting mobility and therefore available drain current, which is sensible mostly in strong

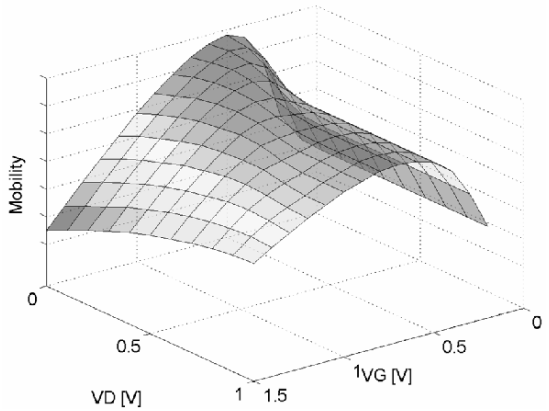


Figure 8. Integral mobility (arbitrary units) versus gate and drain voltage in a long-channel MOST, showing the influence of Coulomb- (low V_G), phonon- and surface roughness (high V_G) scattering, as well as dependence on drain voltage from linear to saturation operation, at $V_S = 0\,\mathrm{V}$.

inversion for velocity saturated conditions. Mobility of the channel carriers is related to drift velocity v_d as,

$$\mu = v_d / E_{II} \tag{29}$$

where $E_{II} = \partial \Psi_s / \partial x$ is the longitudinal field along the channel. The inversion charge linearization vs. surface potential is again conveniently used, $\partial \Psi_s \cong \partial Q_i' / n_q$. A common approach to relate velocity saturation to mobility is the well-known 1st-order hyperbolic model

$$v_d = v_{sat} \frac{E_{II}/E_C}{1 + E_{II}/E_C} \tag{30}$$

where $E_C \cong v_{sat}/\mu_0$ is the critical field for velocity saturation usually considered as a temperature dependent parameter.

The above mobility relationship is easy to handle analytically and is therefore often preferred for simplicity. Theoretically, a 2nd-order velocity-field relationship should be used for electrons. In EKV3.0, a variable-order velocity-field relationship is used as follows,

$$v_d = v_{sat} \frac{E_{II}/E_C}{\sqrt{1 + \dfrac{[2(2-\delta)\cdot(E_{II}/E_C)]^2}{G + |2(2-\delta)\cdot(E_{II}/E_C)|} + (E_{II}/E_C)^2}} \tag{31}$$

where $1 \le \delta \le 2$ is an adjustable parameter defining the order of the velocity-field-relationship and G is a constant. In Figure 9, the variable-order velocity-field relationship is compared with 1st- and 2nd-order relationships.

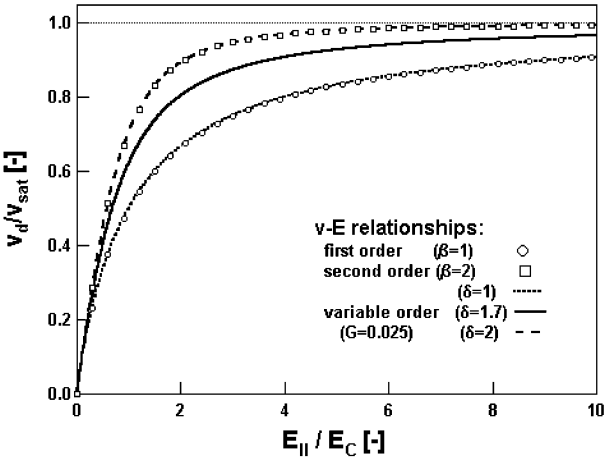


Figure 9. Variable-order velocity-field relationships compared with 1st- and 2nd-order relationships.

The final drain current expression including vertical field and velocity saturation is then,

$$I_D = \frac{2n_q U_T^2 \cdot \bar{\mu}_\perp C'_{OX} \dfrac{W}{L} \left[q_s^2 + q_s - q_d^2 - q_d\right]}{\sqrt{1 + \dfrac{\left[4\varepsilon(2-\delta)\cdot(q_s - q_d)\right]^2}{G + \left|4\varepsilon(2-\delta)\cdot(q_s - q_d)\right|} + \left(2\varepsilon\,(q_s - q_d)\right)^2}} \tag{32}$$

where $\varepsilon = U_T/L \cdot E_C$. Note that this formulation not only introduces a flexible handling of the degree of velocity saturation, it also responds to the need of correctly handling source-drain symmetry at the point $V_D = V_S$.

Output conductance in short-channel transistors in saturation is dominated by channel length modulation mostly in strong inversion. A quasi-two-dimensional approach is used to model the modulation of the channel length ΔL,

$$\Delta L \cong \lambda \cdot \ln\left(1 + \frac{V_{DS} - V_{DSsat}}{L_C \cdot E_C}\right) \quad \text{where} \quad L_C = \sqrt{\frac{\varepsilon_{si} \cdot X_J}{C'_{OX}}} \tag{33}$$

where X_J is the junction depth and λ the adjustable parameter for channel length modulation. The saturation voltage is related to inversion charge densities,

$$V_{DSsat} = U_T \left[2\left(q_s - q'_d\right) + \ln\left(\frac{q_s}{q'_d}\right)\right] \tag{34}$$

$$q'_d \cong q_s + \frac{1}{2}\left(\frac{1}{\varepsilon} + 1 - \sqrt{\frac{1}{4} + \frac{1}{2\varepsilon}\left(\frac{1}{2\varepsilon} + 1 + 2q_s\right)}\right) \tag{35}$$

Further details of handling CLM in the context of EKV3.0 may be found in [10]. Note that mobility expressions are valid for the part of the channel that is not velocity saturated. In the following, the actual channel length is expressed as $L \to L - \Delta L$, and in the previous evaluations, the inversion charge density at the saturation point is considered instead of the at the drain as $q_d \to q'_d$. Precautions are again needed so that no discontinuities are created at $V_D = V_S$.

3.3. Series Resistance

Source and drain series resistance, if handled as distinct elements, cause additional internal nodes and therefore increase simulation time of large circuits. An simple and efficient approach to handle series resistances is to consider their approximate effect on drain current,

$$I_D \cong \frac{I_{D0}}{1 + g_{ms0} \cdot R_S + g_{md0} \cdot R_D} \tag{36}$$

where R_S and R_D are the source and drain resistances, respectively. In the above expression, I_D, g_{ms0} and g_{md0} denote the drain current and source- and drain transconductances evaluated assuming no series resistances are present. The direct relation among transconductances and inversion charge densities $g_{ms} = Y_{\mathrm{Spec}} \cdot q_s$ and $g_{md} = Y_{\mathrm{Spec}} \cdot q_d$ can be conveniently used. Furthermore, since $R \propto \rho_{sh} \cdot L_{\mathrm{dif}} / W$, we obtain,

$$I_D \cong \frac{I_{D0}}{1 + r \cdot q_{s0} + r \cdot q_{d0}} \quad \text{where} \quad r = 2nU_T \mu \rho_{sh} \frac{L_{\mathrm{dif}}}{L} \tag{37}$$

where L_{dif} is the length of the LDD diffusion and ρ_{sh} the sheet resistance.

Besides this simple approach to account internally for series resistance, the model also offers the possibility to add external series resistances, requiring however two additional nodes. The model user has therefore the choice among the more efficient, although less accurate, approach of internally accounting for series resistance, or the external one incurring increased computational effort in large circuits, however providing higher accuracy.

3.4. Short-Channel Effects: DIBL, Charge Sharing, RSCE

In order to control short-channel effects in ultra-deep submicron CMOS, halo or pocket implants are commonly used, as is illustrated schematically in Figure 10. Commonly used techniques are Shallow trench isolation, halo or pocket implants near source and drain to control short-channel effects, salicided gate and junction areas and possibly nitrided oxides to reduce gate current.

Drain induced barrier lowering (DIBL), charge sharing and reverse short-channel effect (RSCE) are the main effects dominating weak inversion operation. In Figure 11 the effect of an increasing longitudinal field on the surface

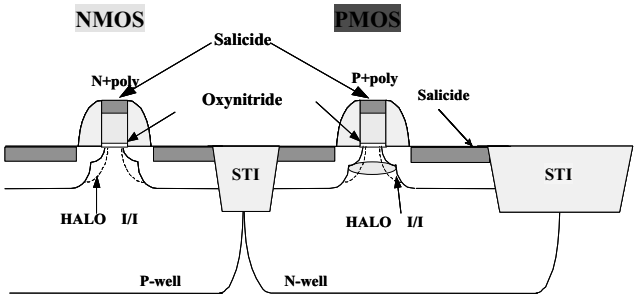


Figure 10. Schematic cross-section of short-channel NMOS and PMOS transistors in an advanced CMOS technology using shallow trench isolation (STI), oxynitride gate oxide, halo implantation near source/drain, and salicided gate and junction areas.

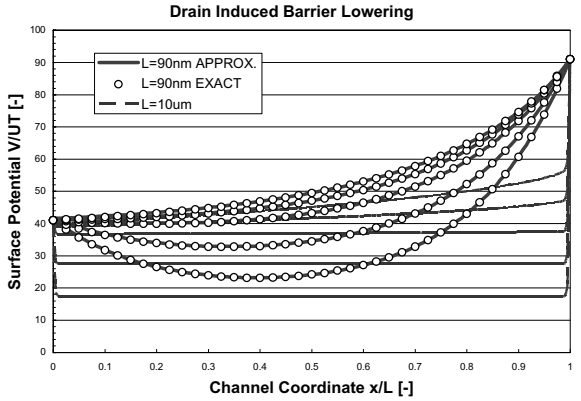


Figure 11. Drain induced barrier lowering effect on surface potential, along the channel, for a long- (lines) and a short-channel (markers and lines) transistors, for fixed V_S and V_D and increasing values of V_P.

potential distribution along the channel can be observed. While the drain voltage has practically no incidence on the surface potential for a long transistor, a short channel transistor is affected significantly by the drain. A quasi-two-dimensional solution of the field distribution near the drain leads to the expression used in EKV3.0.

The DIBL effect is governed by a characteristic length, L_0,

$$L_0 = \eta_D \cdot \sqrt{\frac{\varepsilon_{si} \cdot \gamma}{q \cdot N_{\text{sub}}} \sqrt{\Psi_0}} \tag{38}$$

where $\eta_D \cong 1$ is the main parameter for DIBL. Note the other implicit dependences of L_0 on N_{sub} and C'_{ox}, as well as temperature via Ψ_0 – since the latter

decreases with increased temperature, DIBL tends to have less influence at higher temperature and vice versa.

The estimated difference $\Delta\Psi_S$ of minimum surface potential in the short-channel case, due to DIBL, with respect to the long-channel case, is proportional to,

$$\Delta\Psi_S \propto e^{\left(-\frac{1}{2}\frac{L}{L_0}\right)} \tag{39}$$

and is therefore exponentially dependent on channel length. Note that the exponential is bias-independent and therefore can be evaluated once for each channel length. The $\Delta\Psi_S$ shift is itself approximated by an equivalent shift of the pinch-off voltage $\Delta V_P \approx \Delta\Psi_S$.

The combination of DIBL, charge-sharing and reverse short-channel effect in EKV3.0 gives good results for threshold voltage modelling over channel lengths, as can be seen in Figure 12. Results of threshold voltage modelling

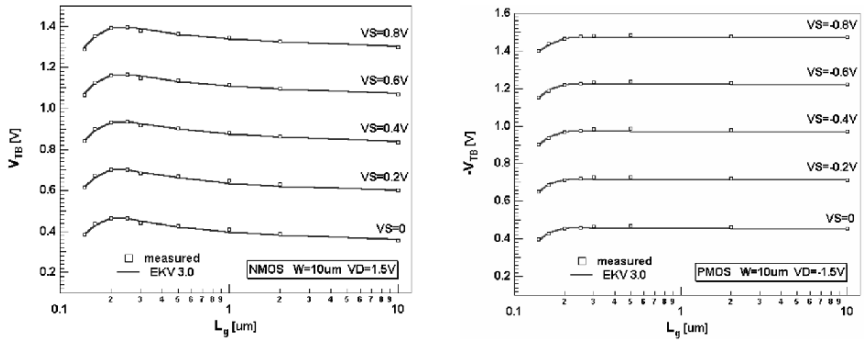


Figure 12. Combined DIBL, charge-sharing and reverse short-channel effect modelling of threshold voltage in 0.14 um CMOS.

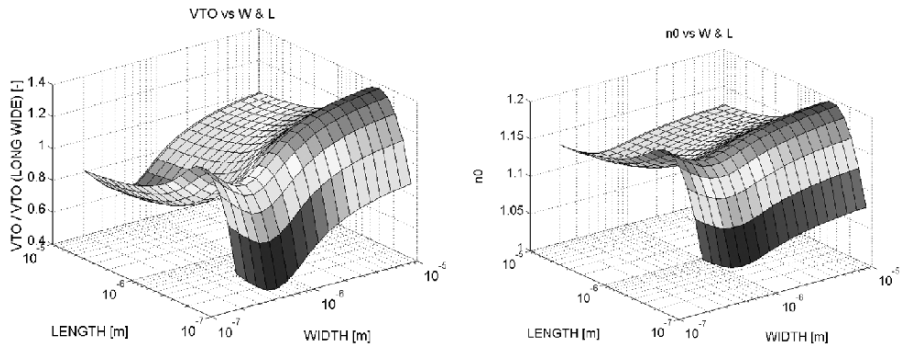


Figure 13. Threshold voltage – relative to long/wide channel – and slope factor n versus channel length and width. Parameter values are realistic for NMOS transistors of an 0.12 um technology.

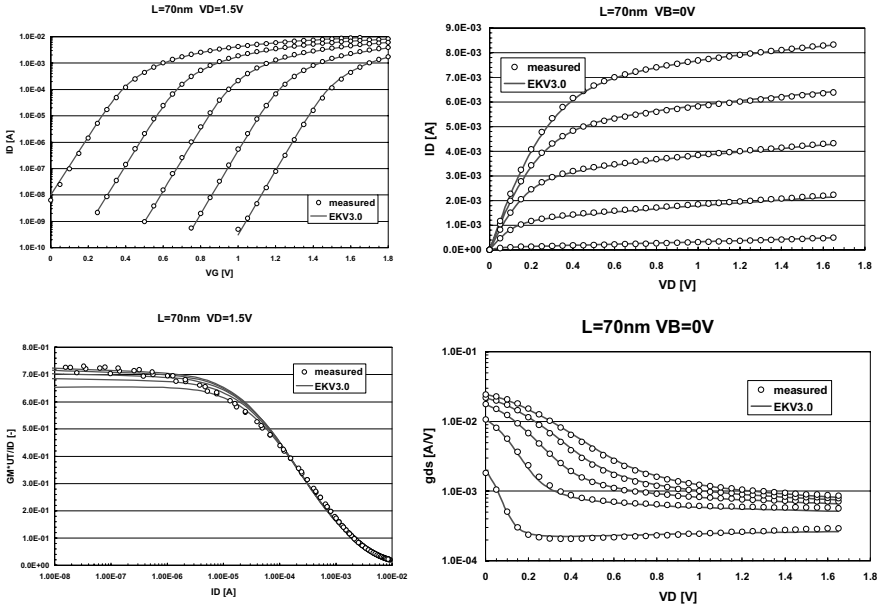


Figure 14. Weak inversion characteristics (left) of a 70 nm n-channel transistor in saturation. The EKV3.0 model provides accurate modelling of weak inversion slope, and transconductance-to-current ratio *versus* normalized drain current. Output characteristics (right) for the same transistor, measurements and modeled with EKV3.0.

for NMOS and PMOS transistors over channel length of an 0.14 um CMOS technology is shown.

The combined short-and narrow-channel effects on threshold voltage and slope factor n are further illustrated in Figure 13. Both characteristics are notably influenced by RSCE and short-channel roll-off mainly due to DIBL and charge sharing. In the width dimension, note the influence of INWE.

In the following, drain current characteristics and related transconductance and output conductance are presented in Figure 14 for an NMOS transistor with effective channel length of 70 nm. These characteristics are all very strongly dependent on DIBL, most notably in weak-moderate inversion.

In order to illustrate these short-channel effects further, Figure 15 shows normalized gate and source transconductance for long- and short-channel transistors of an 0.14 um CMOS technology. Overall the EKV3.0 model represents all characteristics very well. Note that the gate transconductance-to-current ratio in weak/moderate inversion is almost unaffected by channel length, due to a compensating effect among charge sharing (reducing the substrate effect and hence improving weak inversion slope) and DIBL (deteriorating the weak inversion slope).

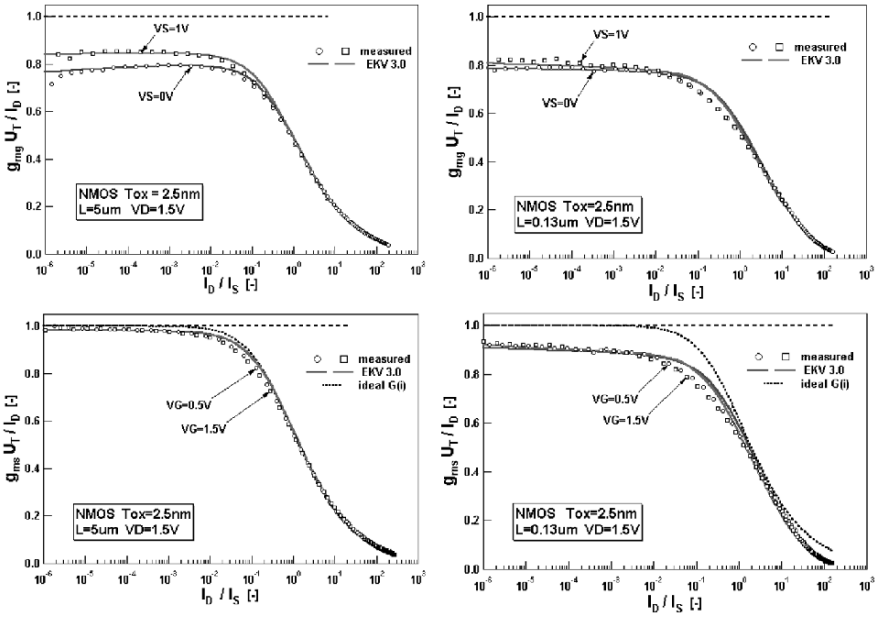


Figure 15. Gate (top) and source (bottom) transconductance to current ratio *versus* normalized drain current, for long-channel (left) and short-channel (right) transistors in 0.14 um CMOS. DIBL effect is responsible for a reduction of g_{ms} in weak inversion for the short-channel transistor.

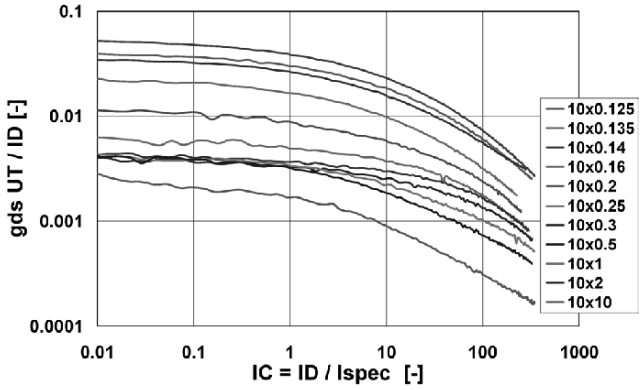


Figure 16. Normalized output conductance-to-current ratio in 0.14 um CMOS. Note that normalized output conductance, instead of improving with longer channels, remains stable or even deteriorates with longer channel lengths (0.3–2 um) in moderate/weak inversion.

Finally, Figure 16 shows normalized output conductance to current ratio versus normalized current for the same technology. It would be expected that normalized output conductance should improve steadily with longer channel lengths. This can be seen not to hold for some intermediate channel lengths,

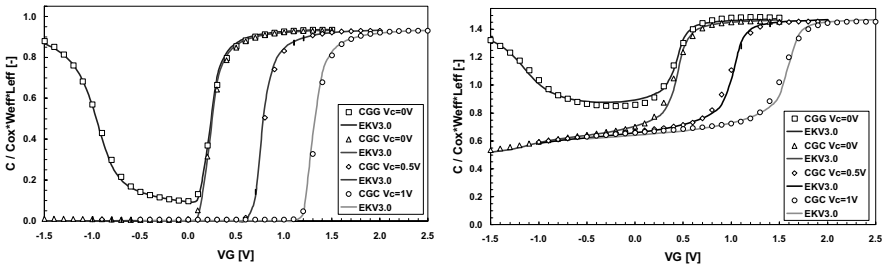


Figure 17. Long- and short-channel NMOS CV characteristics from 0.12 um CMOS, normalized to $C'_{ox}WL$, versus gate voltage, for different channel voltages.

where normalized output conduction to current ratio even deteriorates. This is attributed to the presence of pocket or halo doping implants, degrading output conductance at longer channel lengths.

3.5. Overlap and Fringing Capacitances in Advanced CMOS

Long- and short-channel CV characteristics of an 0.12 um CMOS process are illustrated in Figure 17. Note the correct fitting of all capacitances simultaneously, as well as good fitting of the crucial overlap capacitances in the short-channel characteristics. The latter contribute close to 45% of the total capacitance in (strong) inversion. Overlap and inner fringing capacitances are formulated as charges – preferred over capacitances [25], which are added to the intrinsic channel charges. Note that the overlap capacitances may themselves be affected by polydepletion – just as for MOS varactors in accumulation – when the channel is inverted.

For fringing charge/capacitances, an approach similar to [26] is used. This allows to improve inversion related capacitances in moderate inversion and at the onset of strong inversion significantly as illustrated in Figure 17.

3.6. High-Frequency Application of EKV3.0

One requirement for high-frequency circuit simulation is the consistency among small-signal AC and large-signal transient simulation. While the NQS model presented in the previous section is attractive for its analytical simplicity and its capacity to provide insight in the physics of high-frequency operation of a MOST, it does not provide a solution for transient large-signal simulation.

A convenient approach to solve this problem is the dividing the intrinsic MOS channel into segments as is shown in Figure 18. A number of N channel

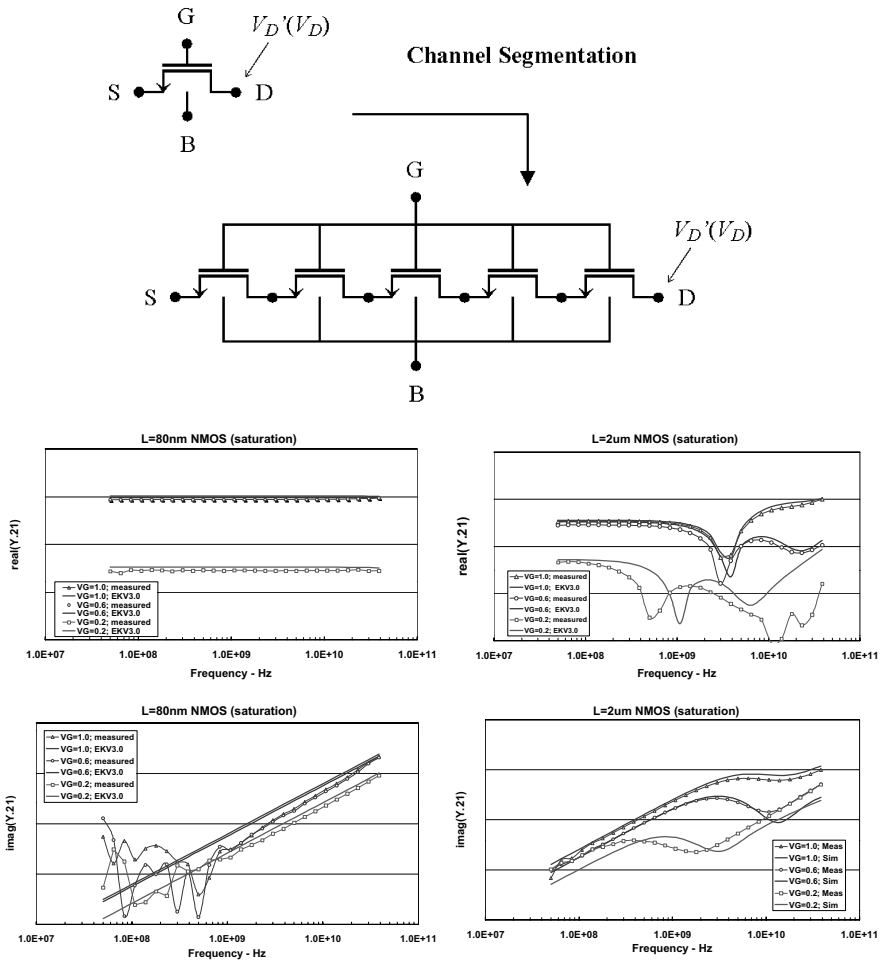


Figure 18. Principle of channel segmentation for consistently handling non-quasistatic (NQS) large-signal and small-signal effects. Real and imaginary part of Y21 for short ($L = 80$ nm, left) and long ($L = 2$ um, right) channel multifinger NMOS transistors operating in saturation and at 3 different gate voltages up to 40 GHz. A qualitatively excellent result is achieved by EKV3.0 with 5 channel segments up to very high frequencies.

segments having an individual length of L/N replace a single transistor with channel length L. This was similarly used in former Philips' models (MM11). A requirement is that the segmented-channel transistor should give the same static and quasistatic response. This can indeed be achieved in the following manner: the mobility of the MOS transistor is calculated for the entire channel, just as if no channel segments existed. Velocity saturation is applied only to the rightmost transistor at the end of the segments chain. For each intermediate

node, only the charge densities need to be evaluated to compute the total charges within each segment.

Figure 18 also presents results on using the above NQS model for small-signal RF modelling with the example of gate transadmittance Y21 in 90 nm CMOS. In the short-channel transistor NQS effects are not visible, while a 2 um transistor shows very significant influence of NQS effects. In the present case, EKV3.0 with 5 channel segments was used. The user may choose the number of segments freely from 1 to 10, according to his/her needs in terms of accuracy.

3.7. Further Aspects Accounted for in EKV3.0

It should be noted that further aspects are included in the EKV3.0 MOS compact model but were not further detailed in the present work. Among these, the following effects should be specially noted:

- Gate tunneling current. The inversion charge linearization principle is extended to account for tunneling through thin oxides. This becomes very significant in 90 nm CMOS technologies and below.
- Substrate current.
- Induced noise in gate and substrate [32].
- Hot-carrier, velocity saturation, mobility and CLM effects on short-channel thermal noise [34].
- Edge conduction effects, resulting from shallow trench isolation.
- Gate and substrate parasitics network for RF application, scaling of parasitics with number of fingers for RF-layout.
- Device matching parameters.
- Temperature effects.

Furthermore, the model is completed with complete sets of extrinsic elements equations, for gate-induced drain/source leakage, diode junction currents and capacitances, according to the BSIM4 model.

In future releases of the model, it is expected that the following effects will be made available:

- Layout-dependent stress effects.
- Matching for gate current.

3.8. Parameters and Principles of Parameter Extraction

Table 2 provides a synoptic overview of the main parameters of EKV3.0. The parameters, written in SPICE syntax, are grouped according to their role (compare with Table 1), and include indicative and/or default values, with

Table 2. List of main parameters (∼ 90) in EKV3.0 with indicative values. Parameters for 2nd order scaling (∼ 30), extrinsic elements (diodes, gate induced drain leakage, series resistance) are not included.

- Flags
  + SIGN = 1
  + TG = -1
- Scale parameters
  + SCALE = 1.0
  + XL = 0.0
  + XW = 0.0
- Cgate parameters
  + COX = 10.0E-3
  + GAMMAG = 6.0
  + AQMA = 0.5
  + AQMI = 0.4
  + ETAQM = 0.75
- Nch. parameters
  + VTO = 200.0E-3
  + PHIF = 450.0E-3
  + GAMMA = 300.0E-3
  + VBI = 1.0
  + XJ = 20.0E-9
  + NO = 1.0
- Mobility
  + KP = 300.0E-6
  + E0 = 1.0E+9
  + E1 = 400.0E+6
  + ETA = 0.5
  + ZC = 1.0E-6
  + THC = 0.0
- Long-ch. gds degr.
  + PDITS = 0.0
  + PDITSD = 0.0
  + PDITSL = 0.0
  + FPROUT =10.0E+6
  + DDITS = 0.3
- Matching par.
  + AVTO = 0.0
  + AKP = 0.0
  + AGAMMA = 0.0
- Vsat & CLM par.
  + UCRIT = 5.0E+6
  + DELTA = 1.5
  + LAMBDA = 0.5
  + ACLM = 0.83
- Geometrical par.
  + DL = 0.0
  + DLC = 0.0
  + WDL = 0.0
  + LL = 0.0
  + LLN = 1.0
  + DW = -10.0E-9
  + DWC = 0.0
  + LDW = 0.0
- Charge sharing
  + LETA0 = 0.0
  + LETA = 1.0
  + LETA2 = 0.0
  + WETA = 1.0
  + NCS = 1.0
- DIBL
  + ETAD = 1.0
  + SIGMAD = 1.0
- RSCE
  + LR = 40.0E-9
  + QLR = 2.5E-3
  + NLR = 100.0E-3
  + FLR = 0.0
- INWE
  + WR = 60.0E-9
  + QWR = 2.0E-3
  + NWR = 50.0E-3
- Series resistance
  + RLX = 50.0E-6
  + LDIF = 100.0E-9
- Overlap & fringing
  + LOV = 10.0E-9
  + GAMMAOV = 2.5
  + VFBOV = 0.0
  + KJF = 0.0
  + CJF = 0.5
- Gate current
  + KG = 30.0E-6
  + XB = 3.1
  + EB = 29.0E+9
  + LOVIG = 20.0E-9
- Substrate current
  + IBA = 100.0E+6
  + IBB = 300.0E+6
  + IBN = 1.0
- Edge device cond.
  + WEDGE = 10.0E-9
  + DGEDGE = 30.0E-3
  + DPEDGE = 20.0E-3
- Temperature par.
  + TNOM = 27.0
  + TCV = 500.0E-6
  + BEX = -1.5
  + TE0EX = 0.0
  + TE1EX = 1.5
  + TETA = 6.0E-3
  + UCEX = 0.8
  + TLAMBDA = 0.0
  + IBBT = 0.0
  + TCVL = 0.0
  + TCVW = 0.0
  + TCVWL = 0.0
- Flicker noise
  + AF = 1.0
  + KF = 1.0E-24
  + EF = 2.0

typical values for an 0.12 um CMOS technology. The reader is cautioned that this list is not exhaustive and that parameter names might slightly differ in the actual computer simulation model.

A few comments on the parameters are in order here. NMOS and PMOS transistors have the same parameter set, with same signs of parameters except for the flag `SIGN=1` which denotes an NMOS transistor, and `SIGN=-1` a PMOS. The type of the gate can be chosen opposite to the channel as usual for

enhancement type transistors with `TG=-1` (e.g. N+ poly for p substrate), while `TG=1` denotes similar type of gate as the channel. The latter may be used e.g. for modeling of MOS varactors, which present polydepletion effect when the channel is accumulated. The scale of parameters (e.g. meter per default, or micrometer) may be chosen with `SCALE`.

In Figure 19, a flowchart for the extraction of the main parameters of EKV3.0 is presented. For simplicity, higher-order effects have been omitted. A set of less than 25 parameters is sufficient to represent current-voltage and capacitance-voltage characteristics over channel length for one type of transistor. It should be noted that such a rough hand parameter extraction can be done even by a non-expert user. Model users of former versions, e.g. EKV2.6, will find many similarities with the formerly existing model. Once a rough set of parameters is obtained including length scaling, refinements need to be done for narrow width effects, combined short/narrow channel effects and temperature. Further details of parameter extraction are described e.g. in [38–41]. If necessary, 2nd order scaling of parameters with geometry can be used to improve the overall fitting over geometry.

3.9. Implementation in Verilog-A, ADMS and Diffusion of C-code

The EKV3.0 model has been fully coded in Verilog-AMS [49] and was tested in several circuit simulators (ADS, ELDO, Spectre). Model implementation

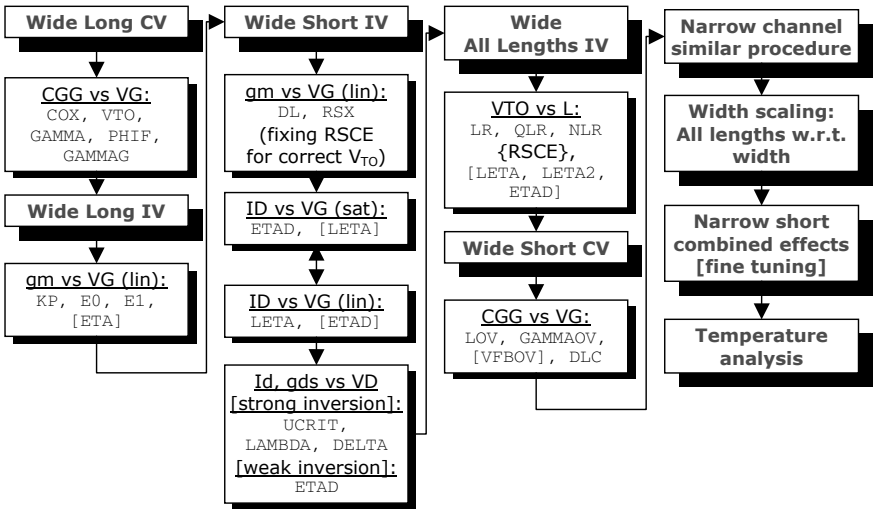


Figure 19. Extraction flowchart showing a possible sequence of basic parameter extraction for EKV3.0 model.

in a Verilog-AMS to C code converter, called ADMS [47], has been completed at the time of writing this chapter. This allows generation of executable C-code for simulators for which XML interfaces in ADMS exist. Notably, SPICE3F5 of UC Berkeley, but also commercial simulators among which Cadence's Spectre, and Synopsys' HSPICE. Further implementations, such as direct C-code implementations in Xpedion's GoldenGate and Mentor Graphics' ELDO are either completed or nearing completion at the time of writing. Therefore, the EKV3.0 code is being made widely available to the community.

4. Conclusions

In summary, this chapter presents aspects of the EKV3.0 model formulation to address modelling of sub-100 nm CMOS. Basics of model formulation, namely the charge-based approach to MOST modelling, have been presented. This provides a consistent approach to model static, quasistatic, non-quaistatic and noise properties of the ideal long-channel MOS transistor. One advantage of the inversion charge linearization model lies in its high analytic versatility. It is therefore particularly suited to advanced analog design. Model extensions to account for high-field effects in advanced CMOS have been outlined and modelling results on various CMOS technologies ranging from 0.25 um to 90 nm CMOS presented.

EKV3.0, with approximately 90 main parameters, accounts for most geometrical, bias and parasitic effects observed in sub-100 nm CMOS technologies. Moreover, the model can be used with a rather small subset of parameters and included effects while leaving others inactive. This facilitates learning as well as teaching, therefore making the model more easily accessible. EKV3.0 is being made available to a wide community for analog/RF circuit design.

Acknowledgments

The authors are grateful to all who have contributed to the initiative of making EKV3.0 possible. In particular, Alain-Serge Porret, Christophe Lallement, Ananda Roy for contributions to model R&D, Wladyslaw Grabinski for Verilog-A support, logistics and management of the website http://legwww.epfl.ch/ekv, Laurent Lemaître for support with ADMS, particular help by Sadayuki Yoshitomi and Joachim Assenmacher, contributions to measurement and parameter extraction by Dimitrios Kazazis and Eleni Kitonaki, as well as partial financial support by Toshiba and Infineon. The authors look forward to constructive feedback by the model users.

References

General

[1] Vittoz, E. "Micropower techniques", *Design of VLSI Circuits for Telecommunication and Signal Processing*, J.E. Franca and Y.P. Tsividis, Eds., Chapter 5, Prentice Hall, **1993**.

[2] Vittoz, E.; Enz, C.; Krummenacher, F. "A basic property of MOS transistors and its circuit implications", *Workshop on Compact Models – 6th Int. Conf. Modeling and Simulation of Microsystems (MSM 2003)* California, USA: San Francisco, **February 2003**, 23–27.

Charge Model/Surface Potential Model Development

[3] Maher, M.A.; Mead, C.A. "A physical charge-controlled model for the MOS transistors", *Advanced Research in VLSI*, P. Losleben, Ed. Cambridge, MA: MIT Press, **1987**.

[4] Iniguez, B.; Moreno, E.G. "A physically based C-finite continuous model for small-geometry MOSFET", *IEEE Trans. Electron Dev.*, **February 1995**, *42(2)*, 283–7.

[5] Enz, C.C.; Krummenacher, F.; Vittoz, E.A. "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications", *J. Analog Int. Circ. Signal Processing*, **1995**, *8*, 83–114.

[6] Cunha, A.I.A.; Schneider, M.C.; Galup-Montoro, C. "An explicit physical model for the long-channel MOS transistor including small-signal parameters", *Solid-State Electron.*, **November 1995**, *38(11)*, 1945–1952.

[7] Cunha, A.I.A.; Gouveia-Filho, O.; Schneider, M.C.; Galup-Montoro, C. "A current-based model of the MOS transistor", *Proc. IEEE Int. Symp. on Circ. & Syst.* (ISCAS'97), **June 1997**, *3*, 1608–1611.

[8] Bucher, M.; Enz, C.; Lallement, C.; Theodoloz, F.; Krummenacher, F. "Scalable GM/I based MOSFET model", *Int. Semicond. Dev. Research Symp. (ISDRS'97)*, Virginia: Charlottesville, **December 1997**, 615–618.

[9] Tsividis, Y. "Operation and modelling of the MOS transistor", 2nd edition, McGraw-Hill, **1999**.

[10] Bucher, M. "Analytical MOS transistor modeling for analog circuit simulation", *Ph.D. Thesis No. 2114 (1999)*, Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland, **2000**.

[11] Enz, C.; Bucher, M.; Porret, A.-S.; Sallese, J.-M.; Krummenacher, F. "The foundations of the EKV MOS transistor charge-based model", *Workshop on Compact Models – 5th Int. Conf. Modeling and Simul. Microsystems (MSM 2002)*, Puerto Rico, USA: San Juan, **April 2002**, 666–669.

[12] Sallese, J.-M.; Bucher, M.; Krummenacher, F.; Fazan, P. "Inversion charge linearization in MOSFET modeling and rigorous derivation of the EKV compact model", *Solid-State Electron.*, **2003**, *47*, 677–683.

[13] Gildenblat, G.; Wang, H.; Chen, T.-L.; Gu, X.; Cai, X. "SP: An advanced surface-potential-based compact MOSFET model", *IEEE J. Solid-State Circuits*, **September 2004**, *39(9)*, 1394–1406.

[14] Watts, J.; McAndrew, C.; Enz, C.; Galup-Montoro, C.; Gildenblat, G.; Hu, C.; van Langevelde, R.; Miura-Mattausch, M.; Rios, R.; Sah, C.-T. "Advanced compact models for MOSFETs", *Workshop on Compact Models – Nanotech 2005*, California, USA: Anaheim, **May 2005**, 9–12.

Polydepletion and Quantum Effects

[15] Sallese, J.-M.; Bucher, M.; Lallement, C. "Improved analytical modelling of polysilicon depletion for CMOS circuit simulation", *Solid-State Electron.*, **June 2000**, *44(6)*, 905–912.

[16] Bucher, M.; Sallese, J.-M.; Lallement, C. "Accounting for quantum effects and polysilicon depletion in an analytical design-oriented MOSFET model", *IEEE Int. Conf. Simul. Semicond. Processes and Dev. (SISPAD 2001)*, D. Tsoukalas and C. Tsamis, Eds., Athens, Greece: Springer, **September 2001**, 296–299, ISBN 3-211-83708-6.

[17] Lallement, C.; Sallese, J.-M.; Bucher, M.; Grabinski, W.; Fazan, P. "Accounting for quantum effects and polysilicon depletion from weak to strong inversion in a charge-based design-oriented MOSFET model", *IEEE Trans. Electron Dev.*, **February 2003**, *50(2)*, 406–417.

Charge/Transcapacitances Modelling

[18] Ward, D.E. "Charge based modeling of capacitance in MOS transistors", Technical Report G201-11, Integrated Circuits Laboratory, Stanford University, **June 1981**.

[19] Bucher, M.; Sallese, J.-M.; Lallement, C.; Grabinski, W.; Enz, C.C.; Krummenacher, F. "Extended charges modelling for deep submicron CMOS", *Int. Semicond. Device Research Symp. (ISDRS'99)*, Virginia: Charlottesville, **December 1999**, 397–400.

[20] Bucher, M.; Enz, C.; Krummenacher, F.; Sallese, J.-M.; Lallement, C.; Porret, A.-S. "The EKV 3.0 MOS transistor compact model: Accounting for deep submicron aspects", (Invited Paper), *Workshop on Compact Models – 5th Int. Conf. Modeling and Simul. Microsystems (MSM 2002)*, Puerto Rico, USA: San Juan, **April 2002**, 670–673.

Mobility Modelling, Low-T MOS Application

[21] Martin, P.; Bucher, M.; Enz, C. "MOSFET modeling and parameter extraction for low temperature analog circuit design", *Journal de Physique IV*, *12*, **2002**, Pr3–51–56, Les Editions de Physique, Les Ulis, France.

[22] Saramad, S.; Anelli, G.; Bucher, M.; Despeisse, M.; Jarron, P.; Pelloux, N.; Rivetti, A. "Modeling of an integrated active feedback preamplifier in a 0.25 µm CMOS technology at cryogenic temperatures", *IEEE Trans. Nucl. Sci.*, **August 2003**, *50(8)*.

[23] Martin, P.; Bucher, M. "Comparison of 0.35 and 0.21 µm CMOS technologies for low temperature operation (77 K–200 K) and Analog Circuit Design", *6th European Workshop on Low Temperature Electronics (WOLTE 6)*, The Netherlands: Noordwijk, **June 2004**, 23–26.

Series Resistance, Overlap Capacitance

[24] Cserveny, S. "Relationship between measured and intrinsic transconductances of MOSFETs", *IEEE Trans. Electron Dev.*, **1990**, *37(11)*, 2413–2414.

[25] Prégaldiny, F.; Lallement, C.; Mathiot, D. "A simple efficient model of parasitic capacitances of deep-submicron LDD MOSFETs", *Solid-State Electron.*, **2002**, *46*, 2191–2198.

[26] Gildenblat, G.; Cai, X.; Chen, T.-L.; Gu, X.; Wang, H. "Reemergence of the surface potential based compact models", *IEDM Tech. Digest*, **2003**, 863–866.

High-Frequency and Noise Modelling of the MOSFET

[27] Enz, C.; Cheng, Y. "MOS transistor modeling for RF IC design", *IEEE Trans. Solid-State Circuits*, **February 2000**, *35(2)*, 186–201.

[28] Porret, A.-S.; Sallese, J.-M.; Enz, C. "A compact non quasi-static extension of a charge-based MOS model", *IEEE Trans. Electron Devices*, **2001**, *48(8)*, 1647–1654.

[29] Scholten, A.; "A large signal non-quasi-static MOS model for RF circuit simulation", *IEEE Int. Conf. Simul. Semicond. Processes and Dev. (SISPAD 2001)*, D. Tsoukalas and C. Tsamis, Eds., Athens, Greece: Springer, **September 2001**, 373–376, ISBN 3-211-83708-6.

[30] Enz, C. "An MOS transistor model for RF IC design valid in all regions of operation", *IEEE Trans. Microwave Theory and Tech.*, **2002**, *50(1)*, 342–359.

[31] Porret, A.-S. "Design of a low-power and low-voltage UHF transceiver integrated in a CMOS process", *Ph.D. Thesis No. 2542 (2002)*, Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland, 2002.

[32] Porret, A.-S.; Enz, C.C. "Non-quasi-static (NQS) thermal noise modeling of the MOS transistor", *IEE Proc. Circuits, Dev. and Syst.*, **2004**, *151(2)*, 155–166.

[33] Bucher, M.; Bazigos, A.; Nastos, N.; Papananos, Y.; Krummenacher, F.; Yoshitomi, S. "Analysis of harmonic distortion in deep submicron CMOS", *Proc. 11th IEEE Int. Conf. Electron., Circ. & Syst. (ICECS 2004)*, 395–398, Tel Aviv, Israel, **December 2004**, 13–15.

[34] Roy, A.S.; Enz, C.C. "Compact modeling of thermal noise in the MOS transistor", *IEEE Trans. Electron Dev.*, **April 2005**, *52(4)*, 611–614.

[35] Yoshitomi, S. "Challenges of compact modeling for deep-submicron RF-CMOS devices", *12th Int. Conf. Mixed Design (MIXDES 2005)*, Krakow, Poland, **June 2005**, 22–25.

Transconductance Analysis

[36] Binkley, D.; Bucher, M.; Foty, D. "Design-oriented characterization of CMOS over the continuum of inversion level and channel length", *Proc. IEEE Int. Conf. Electron., Circ. & Syst. (ICECS'2k)*, Kaslik, Lebanon, **December 2000**, 161–164.

[37] Bucher, M.; Kazazis, D.; Krummenacher, F.; Binkley, D.; Foty, D.; Papananos, Y. "Analysis of transconductances at all levels of inversion in deep submicron CMOS", *Proc. 9th IEEE Conf. Electronics, Circ. & Syst. (ICECS 2002)*, Dubrovnik, Croatia, September 15–18, **2002**, *III*, 1183–1186.

Parameter Extraction

[38] Machado, G.; Enz, C.; Bucher, M. "Estimating key parameters in the EKV MOSFET model for analogue circuit design and simulation", *Proc. IEEE Int. Symp. Circ. & Syst. (ISCAS'95)*, Seattle, Washington, April 30–May 3, **1995**, 1588–1591.

[39] Bucher, M.; Lallement, C.; Enz, C. "An efficient parameter extraction methodology for the EKV MOSFET model", *Proc. IEEE Int. Conf. Microelectronic Test Structures (ICMTS'96)*, Trento, Italy, March 25–28, **1996**, *9*, 145–150.

[40] Lallement, C.; Bucher, M.; Enz, C. "Modelling and characterization of non-uniform substrate doping", *Solid-State Electronics*, **December 1997**, *41(12)*, 1857–1861.

[41] Bazigos, A.; Bucher, M. "The EKV3.0 model code and parameter extraction", *EKV Model Users' Group Meeting and Workshop*, EPFL, Lausanne, Switzerland, November 4–5, **2004**.

Model Development

[42] Bucher, M.; Lallement, C.; Enz, C.; Krummenacher, F. "Accurate MOS modelling for analog circuit simulation using the EKV model", *Proc. IEEE Int. Symp. Circ. & Syst. (ISCAS'96)*, Atlanta, Georgia, **May 1996**, 703–706.

[43] Bucher, M.; Lallement, C.; Enz, C.; Theodoloz, F.; Krummenacher, F. "The EPFL-EKV MOSFET model equations for simulation, version 2.6", *Technical Report*, Electronics Laboratory, EPFL, **June 1997**. [Available Online:] http://legwww.epfl.ch/ekv

[44] Bucher, M.; Enz, C.; Krummenacher, F.; Sallese, J.-M.; Lallement, C.; Porret, A.-S. "The EKV3.0 compact MOS transistor model: Accounting for deep submicron aspects", *Workshop on Compact Models-MSM 2002*, Puerto Rico, **April 2002**, 670–673.

[45] Bucher, M.; Enz, C.; Krummenacher, F.; Sallese, J.-M.; Lallement, C.; Porret, A.-S. "The EKV3.0 compact MOS transistor model: Accounting for deep submicron aspects", *Workshop on Compact Models-MSM 2002*, Puerto Rico, **April 2002**, 670–673.

[46] Bucher, M.; Lallement, C.; Krummenacher, F.; Enz, C. "A MOS transistor model for mixed analog-digital IC design", R. Reis and J. Jess Eds., In *Design of System on a Chip. Devices & Components*, Kluwer Acad. Publ., **2004**, ISBN 1-4020-7928-1.

[47] Bucher, M.; Krummenacher, F.; Bazigos, A. "The EKV3.0 MOSFET model for advanced analog IC design", *EKV Model Users' Group Meeting and Workshop*, EPFL, Lausanne, Switzerland, November 4–5, **2004**.

Verilog-A Modelling

[48] Lemaître, L.; Grabinski, W.; McAndrew, C. "Compact device modeling using Verilog-AMS and ADMS", *Electron Technol. Internet J.*, **2003**, *2(35)*, 1–5, ISSN 0700-9816.

[49] Bazigos, A.; Bucher, M.; Yoshitomi, S. "Benchmarking the EKV3.0 MOSFET model in Verilog-A and 0.14 µm CMOS", *Int. Conf. on Mixed Design (MIXDES 2004)*, Sczcecin, Poland, June 24–26, **2004**, 104–109.

Chapter 4

MODELLING USING HIGH-FREQUENCY MEASUREMENTS

Dominique Schreurs
K.U.Leuven, Div. ESAT-TELEMIC, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
E-mail: Dominique.Schreurs@esat.kuleuven.be

Abstract: This chapter focuses on MOSFET modeling methods that are based on high-frequency measurements directly. Modeling approaches that are based on linear and on non-linear measurements are both explained in detail. The different implementations are illustrated by model examples.

Key words: equivalent circuit model; behavioral model; *S*-parameter measurements; large signal vector measurements; de-embedding.

1. Introduction

The non-linear behavior of MOSFETs has traditionally been described by compact models. This originates from the time that flexible, SPICE compatible, and widely scalable models were required for the silicon based digital designs. Nowadays, the RF performance of silicon CMOS is rapidly increasing and is competing with the III-V compound based devices. As a consequence of which, more and more microwave modeling and design approaches enter the area of silicon analogue circuit design. Whereas most other chapters of this book focus on the latest developments in compact modeling, this and the next two chapters will discuss MOSFET modeling approaches that are based on high-frequency measurements directly. It has to be noted that both the direct high-frequency based and the compact modeling approaches can perform equally well, as recent extensions of compact models are taking special care of high-frequency effects (e.g., gate resistance, substrate network, etc.) [1].

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, 97–119.

This Chapter is structured as follows. First, the theoretical background of two high-frequency modeling approaches is explained: equivalent circuit and behavioral models. A distinction is made between linear and non-linear vector measurements. Subsequently, examples of the different model representations are presented.

2. HF Non-linear Modelling Approaches

2.1. Linear Versus Non-linear Microwave Measurements

The basic principle of linear and non-linear microwave measurements is depicted in Figure 1. In the linear case, a small incident traveling voltage wave a_1 is applied to the device. As response, the device scatters back a scattered traveling voltage wave towards both its port 1 and port 2, which are denoted as b_1 and b_2. If the response has only one spectral component at the same frequency f_0 as the frequency of the excitation signal, then the measurement is linear. In case of microwave measurements, the incident and scattered traveling voltage waves are not measured separately, but only their ratios are characterized. If the non-excited port is loaded by 50 Ohm, we obtain the well-known S-parameters that are being measured by vector network analyzers. The definitions are:

$$\begin{aligned} S_{11} &= \left.\frac{b_1}{a_1}\right|_{a_2=0} \quad S_{12} = \left.\frac{b_1}{a_2}\right|_{a_1=0} \\ S_{21} &= \left.\frac{b_2}{a_1}\right|_{a_2=0} \quad S_{22} = \left.\frac{b_2}{a_2}\right|_{a_1=0} \end{aligned} \tag{1}$$

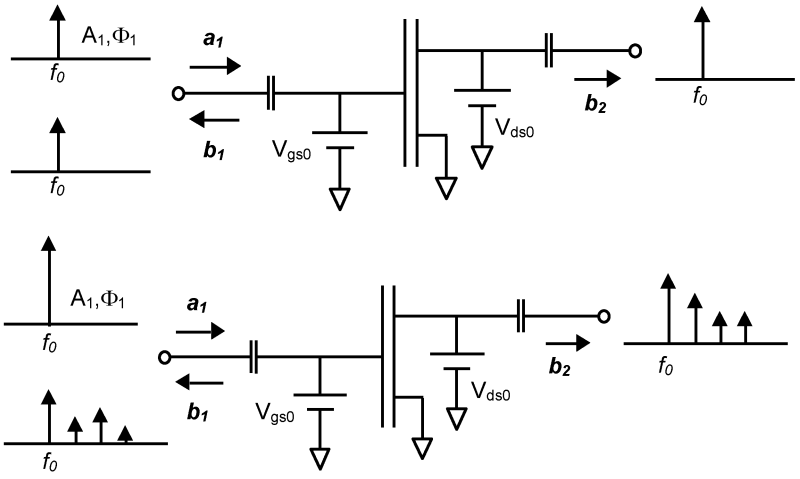


Figure 1. Schematic representation of microwave linear vector measurements (top) and microwave non-linear vector measurements (bottom).

Note that *S*-parameters are complex numbers as both their amplitude and phase are characterized as function of frequency.

In case of non-linear measurements, a larger excitation is applied to the device. In this case, the spectra of the scattered traveling voltage waves have spectral components at not only the excitation, or fundamental, frequency, but also at its harmonics. There exist a variety of instrumentation to measure these non-linear device characteristics. The most complete instrument in the Large-Signal Network Analyzer as it can measure the complex spectra of the incident and scattered traveling voltage waves at both ports simultaneously [2]. On the other hand, a spectrum analyzer only measures the amplitude of the spectral components. Whereas an oscilloscope also considers the phase information, as it measures time domain waveforms, its disadvantages are the difficult calibration and the fact that common instruments are two-channel only (and thus can not measure the four waves simultaneously).

Before proceeding to the different modeling approaches that are based on these measurements, we will first explain why a transistor can behave nonlinearly.

2.2. (Non-)linear Transistor Behavior

Figure 2 shows an example of an NMOS in a linear operation condition. The device is biased at a particular DC bias condition and a small excitation is applied. The device's response can be deduced from its DC characteristics. As the excitation is small, the corresponding AC current swing is linear, which is also reflected by the fact that the output spectrum does not show any harmonics.

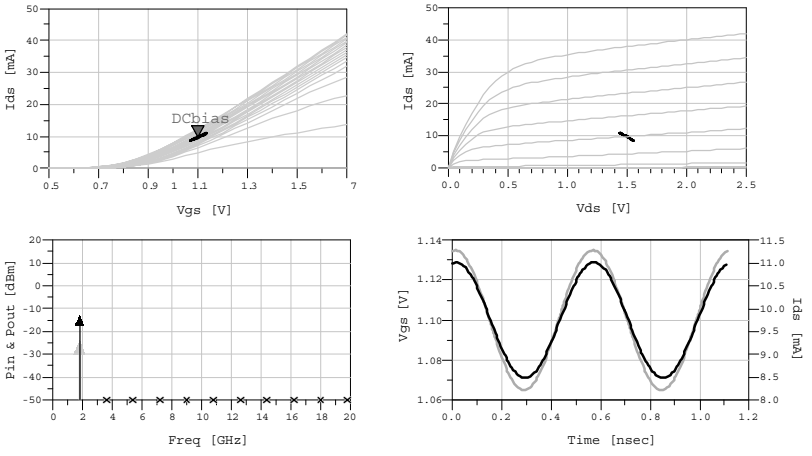


Figure 2. NMOS in linear operation condition ($f_0 = 1.8\,\text{GHz}$, $V_{gs0} = 1.1\,\text{V}$, $V_{ds0} = 1.5\,\text{V}$, load $= 50\,\text{Ohm}$, $P_{in} = -25\,\text{dBm}$).

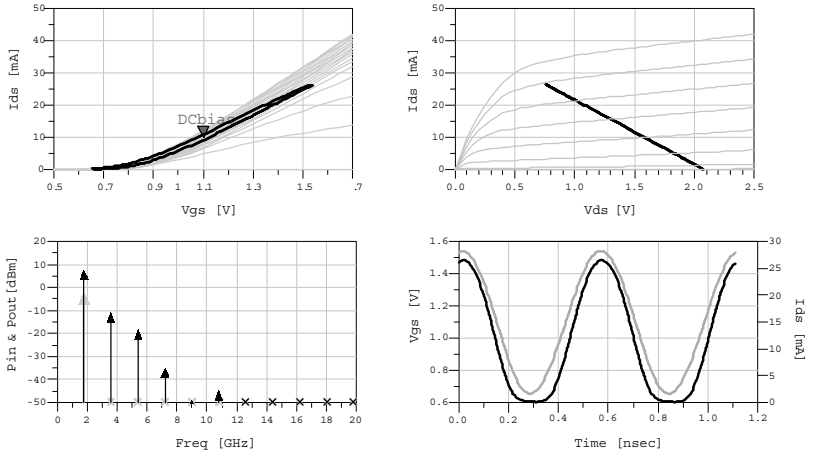


Figure 3. NMOS in non-linear operation condition ($f_0 = 1.8$ GHz, $V_{gs0} = 1.1$ V, $V_{ds0} = 1.5$ V, load $= 50$ Ohm, $P_{in} = -3$ dBm).

If the excitation is large such that the instantaneous current reaches a non-linear part of the DC characteristics, such as the pinch-off region (as illustrated in Figure 3) or the knee region, the response is no longer linear and harmonics are generated.

The purpose of the modeling approaches that will be explained next is to represent this non-linear behavior.

2.3. Linear Measurements Based Models

Figure 4 shows the non-linear quasi-static equivalent circuit of a MOSFET. It is a strongly simplified representation as no extrinsic elements are shown, and also because no second-order effects like dispersion and thermal heating are incorporated. We refer to the next Chapters for a detailed description on the latter, whereas the aim of this Chapter is to explain the general theoretical background.

The non-linear equivalent circuit consists of a charge source at the gate-source terminal, and of a charge and current source at the drain-source terminal. Note that the intrinsic current source I_{dsi} is not the same as the extrinsic DC current. There is no gate-source current source shown, as the gate current in MOSFETs is very small. The terminal voltages are the intrinsic voltages, i.e., after de-embedding the extrinsic elements. By taking partial derivatives, the corresponding small-signal representation can be found. In general terms, the derivative of a charge source is a capacitance, and the derivative of a current

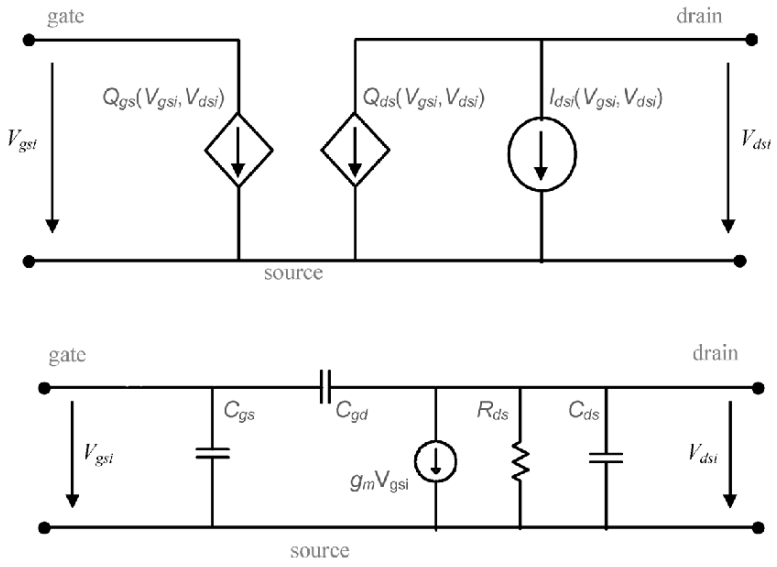


Figure 4. Simplified non-linear (top) and linear (bottom) quasi-static intrinsic equivalent scheme of a MOSFET.

source is a conductance, or:

$$\begin{aligned}
\frac{\partial Q_{gs}\left(V_{gsi}, V_{dsi}\right)}{\partial V_{gsi}} &= C_{11}\left(V_{gsi}, V_{dsi}\right) \quad \frac{\partial Q_{gs}\left(V_{gsi}, V_{dsi}\right)}{\partial V_{dsi}} = C_{12}\left(V_{gsi}, V_{dsi}\right) \\
\frac{\partial Q_{ds}\left(V_{gsi}, V_{dsi}\right)}{\partial V_{gsi}} &= C_{21}\left(V_{gsi}, V_{dsi}\right) \quad \frac{\partial Q_{ds}\left(V_{gsi}, V_{dsi}\right)}{\partial V_{dsi}} = C_{22}\left(V_{gsi}, V_{dsi}\right) \\
\frac{\partial I_{dsi}\left(V_{gsi}, V_{dsi}\right)}{\partial V_{gsi}} &= G_{21}\left(V_{gsi}, V_{dsi}\right) \quad \frac{\partial I_{dsi}\left(V_{gsi}, V_{dsi}\right)}{\partial V_{dsi}} = G_{22}\left(V_{gsi}, V_{dsi}\right)
\end{aligned} \tag{2}$$

When considering the drain-source terminal (a similar analysis can be made at the gate-source terminal), the overall drain-source current is given by the sum of the current source I_{dsi} and the time derivative of the charge source Q_{ds}. When taking the derivative at a particular DC bias point $\left(V_{gsi0,}\, V_{dsi0}\right)$, we obtain:

$$\begin{aligned}
I_{ds}\left(V_{gsi}, V_{dsi}\right) &= I_{dsi}\left(V_{gsi}, V_{dsi}\right) + \frac{dQ_{ds}\left(V_{gsi}, V_{dsi}\right)}{dt} \\
&\quad \downarrow \text{derivative @ } \left(V_{gsi0,}\, V_{dsi0}\right) \\
i_{ds} &= \left(G_{21} + j\omega C_{21}\right) v_{gsi} + \left(G_{22} + j\omega C_{22}\right) v_{dsi} \\
&= Y_{21} v_{gsi} + Y_{22} v_{dsi}
\end{aligned} \tag{3}$$

The $G_{21} + j\omega C_{21}$ and $G_{22} + j\omega C_{22}$ terms can be replaced by the transadmittance Y_{21} and the outputadmittance Y_{22}, respectively.

By measuring the Y-parameters, the reverse operation, i.e., integration, can be performed and the non-linear model is obtained. In practice, it is not possible to measure directly Y-parameters at microwave frequencies as it is hard to realize a perfect short. Therefore, the transistor is characterized by S-parameters at multiple bias points, and subsequently the S-parameters are transformed to Y-parameters.

Finally, the relationship between the well-known small-signal equivalent scheme (shown in Figure 4) and the above analysis has to be clarified. The intrinsic elements have a physical meaning: the capacitances C_{gs} and C_{gd} represent the depletion layer, C_{ds} is the channel capacitance, R_{ds} is the channel resistance, and g_m is the transconductance. The latter models the AC current response at the drain side caused by an AC voltage fluctuation at the gate side. The relationships between the (trans)conductances and (trans)capacitances on one hand, and the intrinsic small-signal equivalent circuit elements on the other hand, are expressed by the following equations:

$$\begin{aligned} C_{11} &= C_{gs} + C_{gd} & C_{12} &= -C_{gd} \\ C_{21} &= -C_{gd} & C_{22} &= C_{ds} + C_{gd} \\ G_{21} &= g_m & G_{22} &= g_{ds} \end{aligned} \tag{4}$$

In summary, the non-linear equivalent circuit based model is mainly based on high-frequency linear vector measurements. This is as opposed to most compact models that are largely based on DC and low-frequency C-V measurements. Also, only measurements on one device are required, in contrary to the need for several devices with various gate length and gate width dimensions in case of compact models. The reason is that compact models are scalable by construction. In case of equivalent circuit models, the dependence on channel width can easily be taken into account, by applying physical knowledge such as that the charges and current are proportional to the width, and that the ohmic source and drain resistances are inversely proportional to the width. Moreover, in analog front-end circuits, mainly devices with minimum channel length are used for highest performance, which reduces the need for channel-length scalable models.

Whereas the small-signal (trans)conductances and (trans)capacitances are the unknowns in the discussed approach, there exist other variants to determine the non-linear model, which will be discussed next.

2.4. Non-linear Measurements Based Models

Figure 5 shows the time domain waveforms of the gate-source voltage and drain-source current at 1 GHz (left) and 30 GHz (right), respectively. When plotting $I_{ds}(t)$ as function of $V_{gs}(t)$, the corresponding bottom plots

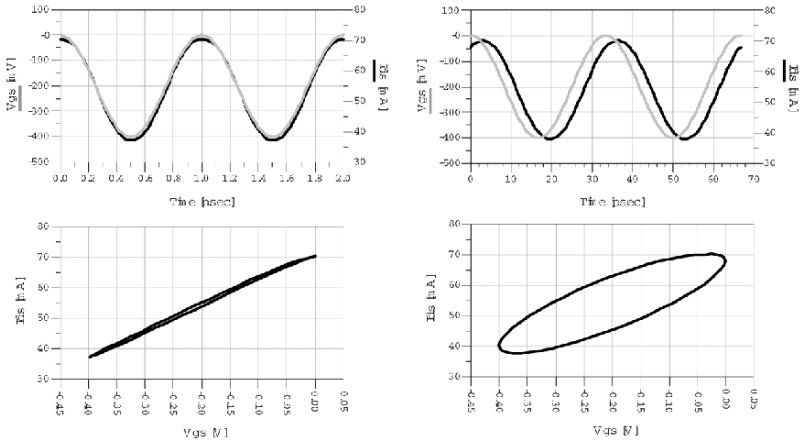


Figure 5. (Top) Time domain waveforms of the gate-source voltage $V_{gs}(t)$ and drain-source current $I_{ds}(t)$ (top), and $I_{ds}(t)$ versus $V_{gs}(t)$ trajectory (bottom). The left hand side plots are at 1 GHz and the right hand side plots are at 30 GHz.

are obtained. It is noticed that $I_{ds}(t)$ is in phase with $V_{gs}(t)$ at low frequencies, whereas there is a time delay at high frequencies. The latter corresponds to a trajectory with hysteresis in the I_{ds}–V_{gs} plane. As I_{ds} is not a single-valued function of V_{gs}, there is at least one independent variable missing.

Eq. (3) can be rewritten as follows:

$$\begin{aligned} I_{ds} &= I_{dsi}\left(V_{gsi}, V_{dsi}\right) + \frac{dQ_{ds}\left(V_{gsi}, V_{dsi}\right)}{dt} \\ \frac{dQ_{ds}\left(V_{gsi}, V_{dsi}\right)}{dt} &= \frac{\partial Q_{ds}\left(V_{gsi}, V_{dsi}\right)}{\partial V_{gsi}}\frac{dV_{gsi}}{dt} \\ &\quad + \frac{\partial Q_{ds}\left(V_{gsi}, V_{dsi}\right)}{\partial V_{dsi}}\frac{dV_{dsi}}{dt} \\ I_{ds} &= I_{dsi}\left(V_{gsi}, V_{dsi}\right) + C_{21}\left(V_{gsi}, V_{dsi}\right)\frac{dV_{gsi}}{dt} + C_{22}\left(V_{gsi}, V_{dsi}\right)\frac{dV_{dsi}}{dt} \\ &= f\left(V_{gsi}, V_{dsi}, \frac{dV_{gsi}}{dt}, \frac{dV_{dsi}}{dt}\right) \end{aligned} \tag{5}$$

This expression confirms that $I_{ds}(t)$ is not only a function of the instantaneous terminal voltages, but also of their first order derivatives.

As Figure 5 represents the type of measurements that can be acquired by a Large-Signal Network Analyzer, several novel modeling approaches become possible. Instead of that the small-signal equivalent circuit elements are the unknowns that need to be determined from the (linear) measurements, several

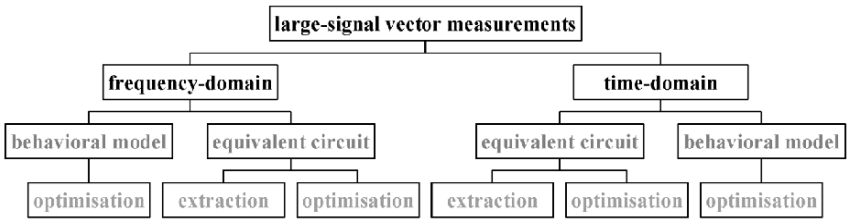


Figure 6. Overview of modeling approaches making use of large-signal vector measurements.

sets of unknowns can be defined that are to be determined from large-signal measurements. The general overview is depicted in Figure 6 [3]. As both the amplitude and phase of the spectral components are characterized, modeling procedures can equally well be developed in time domain and frequency domain [4]. As time domain simulators, such as e.g., SPICE, Spectre, etc., are more common when simulating MOSFET circuits, only the time domain approaches will be addressed in this Chapter.

The two main model representations are the equivalent circuit and the black-box or behavioral model. The former means that unknowns like $I_{dsi}(V_{gsi}, V_{dsi})$, $C_{21}(V_{gsi}, V_{dsi})$, and $C_{22}(V_{gsi}, V_{dsi})$ are determined from the large-signal measurements through either extraction or optimization.

In case of a black-box model, no a-priori physical knowledge about the device is required. The dynamics and thus state variables are determined from the measurements themselves. When considering Eq. (5), it means that the function $f(.)$ combined with the set of independent variables are the unknowns.

The background of this black-box or behavioral model are the well-known state equations:

$$\begin{aligned} \dot{\vec{\mathrm{X}}}(t) &= F_a\left(\vec{\mathrm{X}}(t), \vec{\mathrm{U}}(t)\right) \\ \vec{\mathrm{Y}}(t) &= F_b\left(\vec{\mathrm{X}}(t), \vec{\mathrm{U}}(t)\right) \end{aligned} \tag{6}$$

with $\vec{\mathrm{X}}(t)$ the vector of the state variables, $\vec{\mathrm{U}}(t)$ the vector of the inputs and $\vec{\mathrm{Y}}(t)$ the vector of the outputs. The dot above the symbol is a simplified notation for time derivative. As the inputs in case of MOSFET transistor modeling are usually the voltages and the outputs the currents, the previous set of equations becomes:

$$\begin{aligned} \dot{\vec{\mathrm{X}}}(t) &= F_a\left(\vec{\mathrm{X}}(t), \vec{\mathrm{V}}(t)\right) \\ \vec{\mathrm{I}}(t) &= F_b\left(\vec{\mathrm{X}}(t), \vec{\mathrm{V}}(t)\right) \end{aligned} \tag{7}$$

In practice, a slightly different formulation based on the output expression is used. The generalized expressions for Eq. (5), and similarly for port 1, are:

$$\begin{aligned} I_1(t) &= f_1\left(V_1(t), V_2(t), \dot{V}_1(t), \dot{V}_2(t), \ddot{V}_1(t), \ldots, \dot{I}_1(t), \dot{I}_2(t), \ldots\right) \\ I_2(t) &= f_2\left(V_1(t), V_2(t), \dot{V}_1(t), \dot{V}_2(t), \ddot{V}_1(t), \ldots, \dot{I}_1(t), \dot{I}_2(t), \ldots\right) \end{aligned} \tag{8}$$

As already indicated above, the unknowns in this modeling approach are the functions $f_1(.)$ and $f_2(.)$ combined with the respective sets of independent variables. This model is determined through optimization [5].

The behavioral modeling approach is very general, in the sense that the above formulas cannot only be applied to microwave transistors, but also and even primarily to microwave circuits. The objective in the latter case is to represent the circuit by a lower-order yet accurate dynamical model, in order to speed up system-level simulations. As a circuit design is complex, the loss of physical insight, the major drawback of black-box models, does not really apply. On the other hand, it is important to have a link between the non-linear model and the internal physical operation in case of transistors. For this reason, the modeling examples in next Section are all equivalent circuit related.

3. HF Non-linear Model Examples

Three non-linear MOSFET models will be presented: a look-up table model, which is a particular implementation of a small-signal measurements based equivalent circuit model, and two approaches that determine the equivalent circuit from large-signal measurements directly, being through optimization and extraction, respectively.

3.1. Non-linear Look-up Table Model

A look-up table model is a particular way of representing a large-signal equivalent circuit model. It means that the device's state functions get tabulated as function of bias, and thus that they are not represented by an empirical function. Empirical models are addressed in large detail in the next two chapters.

The basic steps of the modeling procedure are as follows [1]: we first focus on the extraction of the small-signal equivalent circuit. Subsequently, we proceed to the corresponding non-linear model that is obtained from the small-signal model through integration, and we finally evaluate its accuracy by comparing simulations to measurements.

3.1.1. De-embedding of access transmission lines

The MOSFET belongs to the field-effect device family, which means that the small-signal and large-signal equivalent circuit topologies are similar to those of MESFETs and HEMTs [6, 7]. Therefore, an important difference in the non-linear modeling procedures for different FET types lies in the determination of the access elements. We make a clear distinction between, on one hand, the bonding pads and access transmission lines and, on the other hand, the extrinsic, parasitic device elements, because the aim is to construct a model for a device as how it would be inserted in an actual circuit design.

The first step in the modeling procedure is hence the de-embedding of the access transmission lines. Due to the low resistivity of standard silicon substrates, these transmission lines are strongly dispersive. Since EM simulators are time-consuming and often do not provide the required accuracy for this kind of structures, on-wafer calibration is often preferred. The drawback however is that special passive structures need to be foreseen on the mask set.

A possible approach to move the reference plane from the probe tips to the device plane is the three-step de-embedding method [8]. It assumes that the pads and access transmission lines can be represented by a network as shown in Figure 7. Knowing that an elementary section of transmission line can be represented by the parallel connection of a capacitance and conductance in series with an inductance and resistance, it can be understood that a good model for a short section of dispersive transmission line is a parallel complex conductance (e.g., G_1) in series with a complex impedance (e.g., Z_1).

To be able to determine the unknown impedances $Z_{1,2,3}$ and admittances $G_{1,2,3}$, it is necessary to have four dummy passive structures on the wafer: an open, a through, a short-circuit between the gate and source access transmission lines ('short 1'), and a short-circuit between the drain and source access

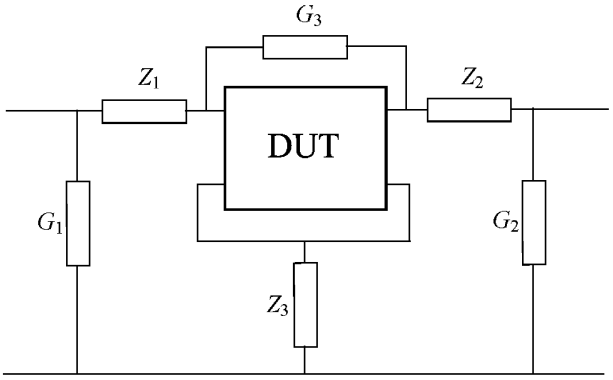


Figure 7. Schematic representing the pads and access transmission lines.

transmission lines ('short 2'). From the S-parameter measurements of these dummy structures, the unknowns can be calculated:

$$\begin{aligned}
G_1 &= Y_{11_{\text{open}}} + Y_{12_{\text{open}}} & Z_1 &= \frac{1}{2}\left\{\frac{-1}{Y_{12_{\text{through}}}} + \frac{1}{Y_{11_{\text{short1}}} - G_1} + \frac{-1}{Y_{22_{\text{short2}}} - G_2}\right\} \\
G_2 &= Y_{22_{\text{open}}} + Y_{12_{\text{open}}} & Z_2 &= \frac{1}{2}\left\{\frac{-1}{Y_{12_{\text{through}}}} + \frac{-1}{Y_{11_{\text{short1}}} - G_1} + \frac{1}{Y_{22_{\text{short2}}} - G_2}\right\} \\
G_3 &= \left(-1/Y_{12_{\text{open}}} + 1/Y_{12_{\text{through}}}\right)^{-1} & Z_3 &= \frac{1}{2}\left\{\frac{1}{Y_{12_{\text{through}}}} + \frac{1}{Y_{11_{\text{short1}}} - G_1} + \frac{1}{Y_{22_{\text{short2}}} - G_2}\right\}
\end{aligned} \tag{9}$$

Next, the contribution by the impedances $Z_{1,2,3}$ and admittances $G_{1,2,3}$ can be de-embedded by a sequence of Y-, Z-, and S-parameter transformations.

In case of non-linear measurements, the pads and access transmission lines need to be de-embedded as well. The formulas can no longer be in terms of S-parameters, as S-parameters are linear by definition. Therefore, the corresponding equations are expressed in terms of currents and voltages (Figure 8):

$$\begin{aligned}
i'_{gs} &= i_{gs} - v_{gs}G_1 \\
i'_{ds} &= i_{ds} - v_{ds}G_2 \\
v'_{gs} &= v_{gs} - i'_{gs}Z_1 \\
v'_{ds} &= v_{ds} - i'_{ds}Z_2 \\
v_{gs,DUT} &= v'_{gs} - \left(i'_{gs} + i'_{ds}\right)Z_3 \\
v_{ds,DUT} &= v'_{ds} - \left(i'_{gs} + i'_{ds}\right)Z_3 \\
i_{gs,DUT} &= i'_{gs} - \left(v_{gs,DUT} - v_{ds,DUT}\right)G_3 \\
i_{ds,DUT} &= i'_{ds} - \left(v_{ds,DUT} - v_{gs,DUT}\right)G_3
\end{aligned} \tag{10}$$

3.1.2. Extraction of extrinsic elements

After de-embedding the pads and access transmission lines, the extrinsic elements have to be determined. A widespread procedure for MESFETs is the cold method ($V_{ds0} = 0\,\text{V}$) developed by Dambrine *et al*. [9]. This method requires a non-negligible gate-current to extract the extrinsic resistances and inductances. Because HEMTs already start to degrade at the required gate

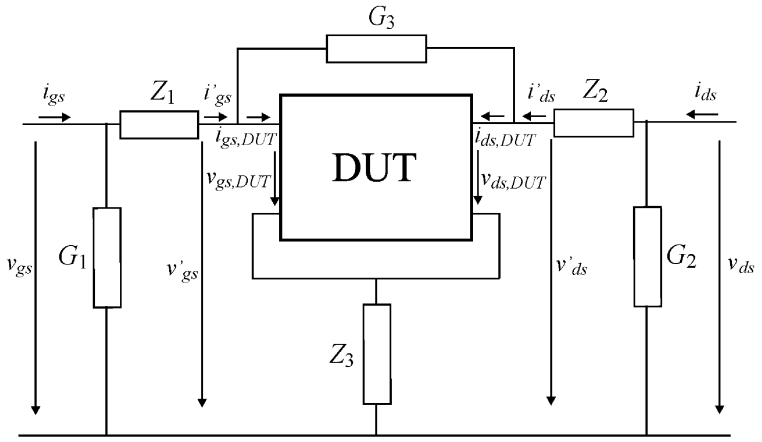


Figure 8. Three-step de-embedding method in case of non-linear measurements.

current level, a modified cold method to overcome this condition has been developed [10]. As the gate current level of MOSFETs is very small, the modified cold method can be applied. In this way, values for the extrinsic resistances are obtained from the following equations:

$$\begin{aligned} \mathrm{Re}\,(Z_{11}) &= R_g + \frac{R_i}{2} + \frac{R_{ch}}{4} + R_s \\ \mathrm{Re}\,(Z_{12}) &= \mathrm{Re}\,(Z_{21}) = \frac{R_{ch}}{2} + R_s \\ \mathrm{Re}\,(Z_{22}) &= R_d + R_{ch} + R_s \end{aligned} \tag{11}$$

with R_{ch} the channel resistance and R_i the non-quasi-static resistance in series with C_{gs}.

Since both R_{ch} and R_i are inversely proportional to $V_{gs} - V_T$ [11], with V_T the linearly extrapolated threshold voltage, the series resistances R_s, R_d and R_g can be extracted subsequently from the plots of Z_{12}, Z_{22}, and Z_{11} versus $1/(V_{gs} - V_T)$, respectively. The intercepts yield values for R_s, R_d and R_g.

The values of the extrinsic capacitances and inductances extracted using the cold method are almost negligible, which is an implication of the three-step de-embedding method: it is difficult to clearly separate the contribution of the access parts and the contribution of the device parasitics.

3.1.3. Extraction of intrinsic elements

After de-embedding the extrinsic part, the bias-dependent intrinsic elements can be extracted. To have a consistent transition between the small-signal and large-signal equivalent schemes, we use the representation with the

transcapacitances [6], as explained in Section 2 of this Chapter. The set of extraction equations is repeated here for completeness:

$$\begin{aligned} C_{11} &= \frac{\mathrm{Im}\,(Y_{11})}{\omega} = C_{gs} + C_{gd} & C_{12} &= \frac{\mathrm{Im}\,(Y_{12})}{\omega} = -C_{gd} \\ C_{21} &= \frac{\mathrm{Im}\,(Y_{21})}{\omega} = -C_{gd} & C_{22} &= \frac{\mathrm{Im}\,(Y_{22})}{\omega} = C_{ds} + C_{gd} \\ G_{21} &= \mathrm{Re}\,(Y_{21}) = g_m & G_{22} &= \mathrm{Re}\,(Y_{22}) = g_{ds} \end{aligned} \tag{12}$$

It has to be reminded that the intrinsic scheme adopted here (Figure 4) is a simplified representation. To have a full-blown model, non-quasi-static effects and other physical phenomena such as dispersion and thermal heating should be accounted for. Examples of such equivalent circuit based models will be discussed in the next two chapters.

To determine the bias-dependency of the intrinsic elements, *S*-parameter measurements are performed at multiple bias points. The bias steps depend on the application. If the model is aimed for use in the saturation region, the typical step sizes are 50 mV for V_{gs} and 100 mV for V_{ds}. In case of a cold application, such as a resistive mixer, a denser (e.g., 20 mV) and also negative V_{ds} grid is required.

3.1.4. Non-linear look-up table model

The final step in the non-linear modeling procedure is the integration of the bias-dependent intrinsic elements towards the corresponding terminal voltages:

$$\begin{aligned} Q_{gs}\left(V_{gsi}, V_{dsi}\right) &= \int_{V_{gsi0}}^{V_{gsi}} C_{11}\left(V, V_{dsi0}\right)dV + \int_{V_{dsi0}}^{V_{dsi}} C_{12}\left(V_{gsi}, V\right)dV \\ Q_{ds}\left(V_{gsi}, V_{dsi}\right) &= \int_{V_{gsi0}}^{V_{gsi}} C_{21}\left(V, V_{dsi0}\right)dV + \int_{V_{dsi0}}^{V_{dsi}} C_{22}\left(V_{gsi}, V\right)dV \\ I_{dsi}\left(V_{gsi}, V_{dsi}\right) &= I_{dsi}\left(V_{gsi0}, V_{dsi0}\right) + \int_{V_{gsi0}}^{V_{gsi}} G_{21}\left(V, V_{dsi0}\right)dV \\ &\quad + \int_{V_{dsi0}}^{V_{dsi}} G_{22}\left(V_{gsi}, V\right)dV \end{aligned} \tag{13}$$

where $\left(V_{gsi0}, V_{dsi0}\right)$ denotes the starting point for the integration, and the value of the current at this bias condition, $I_{dsi}\left(V_{gsi0}, V_{dsi0}\right)$, is the corresponding integration constant. The results should be integration-path independent to ensure charge conservation [12].

The obtained Q_{gs}, Q_{ds}, and I_{ds} are two-dimensional tables as function of bias. In case of empirical models, there is one additional step in the modeling procedure: an empirical function is fitted to the values within the tables. If

the circuit simulator however supports that the tables can be imported and evaluated during the simulations, the additional step of function fitting can be omitted. This is what is called a 'look-up table model'. There is however a drawback: during a simulation, and especially during a harmonic-balance simulation, the table may get evaluated outside its range, which often initiates convergence problems. In case of empirical models, the analytical expressions can be conceived in such a way that the asymptotes of the functions are smooth.

The non-linear model is completed by adding the lumped components that represent the bias-independent extrinsic elements. In the next Section, experimental results are shown.

3.1.5. Model verification

The developed procedure has been applied to a 36-finger nMOSFET with a channel length of 0.18 μm and a total gate width of 146 μm [13]. A complete model verification consists of a check against DC, *S*-parameter, and large-signal measurements. Only the latter will be illustrated in the following.

Non-linear models can be fully verified only by large-signal *vecto*r measurements, meaning that not only the accuracy of the simulated magnitude but also that of the phase of the spectral components of voltages and currents (or voltage waves) can be evaluated. As clarified in Section 3.1.1, the reference plane of the large-signal measurements has to be shifted to the device plane by applying the three-step de-embedding method.

Figure 9 shows the excellent agreement between the measured and simulated first three harmonics of the output power.

An advantage of the measurement set-up is that models also can be verified under more complex excitations [1, 14]. This is illustrated in Figure 10, which shows the I_{gs} and I_{ds} time domain waveforms under a two-tone excitation.

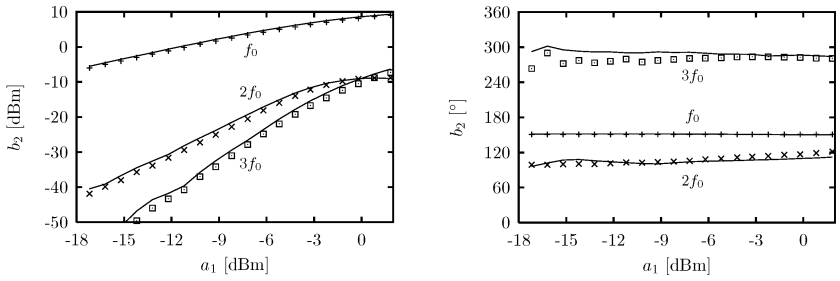


Figure 9. Measured (symbols) and simulated (solid line) magnitude (left) and phase (right) of the first three harmonics of the output power of a 0.18 μm × 146 μm nMOSFET ($V_{gs0} = 0.6$ V, $V_{ds0} = 1.2$ V, $f_0 = 3.6$ GHz).

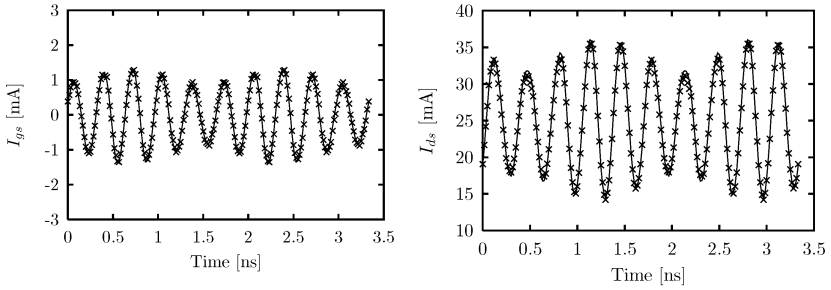


Figure 10. Measured (x) and simulated (solid line) time domain waveforms of the terminal currents of a 0.18 µm × 146 µm nMOSFET ($V_{gs0} = 0.9$ V, $V_{ds0} = 1.2$ V, $f_1 = 3$ GHz, $f_2 = 3.6$ GHz, $a_1 = -10$ dBm, $a_2 = -7$ dBm, $\phi(a_2) - \phi(a_1) = 37^\circ$).

A single-tone signal at 3 GHz is applied at port 1, while another single-tone signal at 3.6 GHz is applied at port 2. Such an excitation can be of interest for mixer applications. The phase difference between the two-signals is arbitrarily chosen and equals 37 deg. In this case as well, an excellent agreement between simulations and measurements can be observed.

In summary, the construction of a large-signal look-up table model for MOSFETs has been explained step by step. The basic procedure is similar to the way that such models are constructed for III–V compound devices, except for the fact that special attention has to be paid to the accurate de-embedding of the dispersive silicon access transmission lines. The non-linear model, being verified by large signal vector measurements, is shown to be accurate in both frequency and time domain.

3.2. Large-signal Measurements Based Model Through Optimization

Next, we will discuss two modeling methods that determine the MOSFET large-signal equivalent circuit *directly* from high-frequency large-signal vector measurements and thus eliminate the small-signal detour [15]. This Section will focus on the optimization-based approach, while the direct extraction approach will be covered in the next Section.

3.2.1. Model parameter estimation procedure

The model adopted is the intrinsic large-signal equivalent circuit as shown in Figure 4. It assumes that the pads and access transmission lines, as well as the bias-independent extrinsic elements have already been calculated, e.g.,

using the method as described in the previous Section, and de-embedded. The objective is to determine the charge sources and the current source from large-signal vector measurements directly.

As explained in Section 2 of this Chapter, the model equations can be written as follows:

$$\begin{aligned} I_{gs} &= C_{11}\left(V_{gsi}, V_{dsi}\right)\frac{dV_{gsi}}{dt} + C_{12}\left(V_{gsi}, V_{dsi}\right)\frac{dV_{dsi}}{dt} \\ I_{ds} &= I_{dsi}\left(V_{gsi}, V_{dsi}\right) + C_{21}\left(V_{gsi}, V_{dsi}\right)\frac{dV_{gsi}}{dt} + C_{22}\left(V_{gsi}, V_{dsi}\right)\frac{dV_{dsi}}{dt} \end{aligned} \tag{14}$$

The unknowns to be determined are the intrinsic drain-source current source and the (trans)capacitances. As the method described hereafter is not suitable for look-up table models, only the case whereby the equivalent circuit elements are represented by analytical functions is considered. Several classes of analytical functions can be distinguished, such as empirical models and artificial neural networks. Both are characterized by a number of parameters of which the values can be estimated by optimization. Whereas these model parameters are classically determined by optimizing the analytical functions towards the DC measured drain-source current (after transformation to the intrinsic bias plane) and the *S*-parameter measurement based capacitances, it is sufficient to fit the model parameters to large-signal vector measurements only [16]. The reason is that this type of measurements contains all necessary information, i.e., both the amplitude and the phase of the spectral components of the incident and scattered traveling voltage waves at both device terminals.

The procedure starts by performing a number of large-signal vector measurements, called 'experiments', where it is possible to sweep any degree of freedom, like input power, excitation frequency, DC bias, load impedance, etc., but depending on the envisaged application one can focus on particular experiments. Subsequently, the parameters of the non-linear model are estimated during one global harmonic balance optimization process in which all experiments are combined. The advantage of this approach is that only one type of measurements, i.e., large-signal vector measurements, and only one type of simulation, i.e., harmonic balance analysis, are needed. It is possible to include "DC"- or "*S*-parameter"-like information by choosing the appropriate operation conditions, e.g., a low input power, when performing the large-signal vector measurements. As the instrument captures the data at both device ports simultaneously, the analytical functions for all equivalent circuit elements can be optimized at once.

3.2.2. Empirical model

This optimization technique can be applied to both artificial neural networks and empirical models. As empirical non-linear model, we choose a simple

version of the Chalmers model [17] (the extended version will be covered in a separate Chapter). The expression for the drain-source current is:

$$\begin{aligned} I_{dsi} &= I_{g\max}\left(1+\tanh\left(\Psi\right)\right)\left(1+\lambda V_{dsi}\right)\tanh\left(\alpha V_{dsi}\right) \\ \Psi &= P_1\left(V_{gsi}-V_{g\max}\right)+P_2\left(V_{gsi}-V_{g\max}\right)^2 \\ &\quad + P_3\left(V_{gsi}-V_{g\max}\right)^3+\cdots \\ V_{g\max} &= V_{g\max 0}+\gamma V_{dsi} \end{aligned} \tag{15}$$

The unknown model parameters to be determined during the optimization are $I_{g\max}$, λ, α, $V_{g\max 0}$, γ, and P_1 to P_3.

Measurement data have been collected on an nMOSFET [13]. The device is operated in class A, while the input power is swept. All the model parameters are simultaneously optimized towards these large-signal vector measurements.

Figure 11 compares the measured and simulated $I_{gs}(t)$ versus $V_{gs}(t)$ and $I_{ds}(t)$ versus $V_{gs}(t)$ at a high input power. This Figure clearly indicates an excellent agreement and hence high model accuracy.

A general drawback with optimization-based models is that the extrapolation capabilities are limited. In other words, if the above model were to be used at significantly distinct operation conditions than the ones included in the optimization, the accuracy will be compromised and will depend on how well the analytical function represents the overall transistor's behavior.

3.2.3. Artificial neural network

As second example of the optimization based method, we consider an Artificial Neural Network (ANN) [18]. A common representation is the three-layer perceptron, represented in Figure 12. The relationship between the N_x input

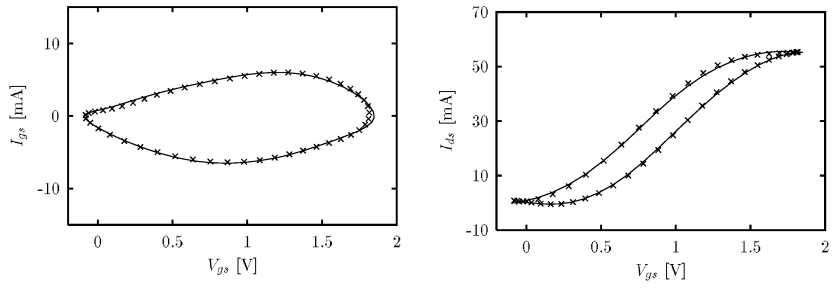


Figure 11. Comparison of the measured (x) and Chalmers modeled (solid line) $I_{gs}(t)$ (left) and $I_{ds}(t)$ (right) versus $V_{gs}(t)$ for a 0.18 µm × 146 µm nMOSFET ($V_{gs0} = 0.9$ V, $V_{ds0} = 1.8$ V, $f_0 = 3.6$ GHz, $P_{in} = 3.8$ dBm).

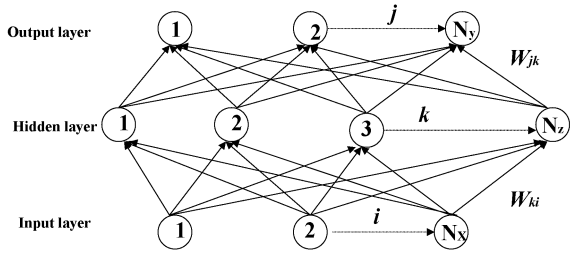


Figure 12. Schematic representation of a three-layer artificial neural network.

and N_y output variables is modeled by means of one hidden layer that has N_z hidden neurons.

The inputs to the hidden layer are the γ_k, calculated from the input variables by:

$$\gamma_k = \left(\sum_{i=1}^{N_x} x_i w_{ki} \right) + \theta_k \tag{16}$$

with $k = 1, 2, \ldots, N_z$, w_{ki} being the weighting factors and θ_k the bias term.

The base function of the hidden layer $f(\zeta)$ is called the 'activation function'. Common activation functions are the sigmoid function $f(\zeta) = 1/(1 + e^{-\zeta})$ and $f(\zeta) = \tanh(\zeta)$. Consequently, the output from the kth neuron of the hidden layer is z_k:

$$z_k = f(\gamma_k) \tag{17}$$

Finally, the output of the jth neuron in the output layer is:

$$y_j = \left(\sum_{k=1}^{N_z} z_k w_{jk} \right) + \eta_j \tag{18}$$

with $j = 1, 2, \ldots, N_y$, w_{jk} being the weighting factors and η_j the bias term.

The training process is in fact an optimization problem to find the best values for $w_{ki}, \theta_k, w_{jk}, \eta_j$ to minimize the objective function, which is the difference between the output from the ANN and the training data. The latter are the data collected from the large-signal vector measurements.

As it is the purpose to keep the link with the device physics and thus with the equivalent circuit, the transistor is not represented by one global artificial neural network (as it would be in case of behavioral models [19]), but each large-signal equivalent circuit element gets represent by an ANN.

In the considered example of the nMOSFET, I_{dsi} got represented by an ANN with five hidden nodes, while three hidden nodes are sufficient to model

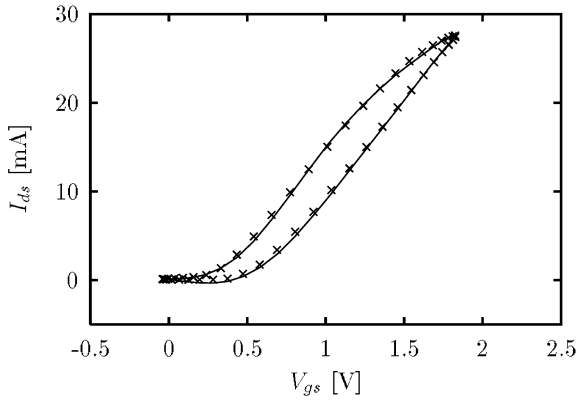


Figure 13. Comparison of the measured (x) and artificial neural network modeled (solid line) $I_{ds}(t)$ versus $V_{gs}(t)$ of a 0.25μm × 50 μm nMOSFET ($V_{gs0} = 0.9\,\text{V}$, $V_{ds0} = 1.8\,\text{V}$, $f_0 = 0.9\,\text{GHz}$, $P_{\text{in}} = 3.4\,\text{dBm}$).

Q_{gs} and Q_{ds}. Note that this number may have to be increased if the targeted operation range is larger.

This ANN modeling approach was applied to an nMOSFET operated in class A. Figure 13 shows the excellent agreement between the measured and modeled $I_{ds}(t)$ versus $V_{gs}(t)$.

As already touched upon above, models determined by optimization procedures often have limited extrapolation capabilities. The choice of a particular empirical expression or the number of hidden nodes and layers is often a compromise between simplicity and accuracy. Therefore, this large-signal measurements based optimization method is preferably used when a device has to be modeled for a well-defined application. In such cases, the degrees of freedom of the measurement set-up can be engineered in such a way that the large-signal vector measurements cover the possible operation conditions that instantaneously can be reached when the device is inserted in that particular circuit application.

3.3. Large-signal Measurements Based Model Through Extraction

The complete information of large-signal vector measurements also allows the direct extraction of the large-signal equivalent scheme. This can be understood from the $I_{ds}(t)$ versus $V_{gs}(t)$ trajectory in Figure 5. As explained in Section two of this Chapter, the shown hysteresis is caused by the capacitive effects, which demonstrates that the contributions of both charge and current sources are clearly visible in the large-signal vector measurements.

The procedure is based on the same set of model equations as before:

$$\begin{aligned} I_{gs} &= C_{11}\left(V_{gsi}, V_{dsi}\right)\frac{dV_{gsi}}{dt} + C_{12}\left(V_{gsi}, V_{dsi}\right)\frac{dV_{dsi}}{dt} \\ I_{ds} &= I_{dsi}\left(V_{gsi}, V_{dsi}\right) + C_{21}\left(V_{gsi}, V_{dsi}\right)\frac{dV_{gsi}}{dt} + C_{22}\left(V_{gsi}, V_{dsi}\right)\frac{dV_{dsi}}{dt} \end{aligned} \tag{19}$$

In this case, no empirical or artificial neural network representation is assumed, but the values of the five unknowns are calculated directly from the large-signal measurements [20]. This is what is called 'extraction' as opposed to 'optimization'.

There are two equations and five unknowns. When only a single-tone CW excitation is applied at the gate, this set of equations can be solved by performing three measurements at three different pulsations. The condition is to have three time points t_1, t_2, and t_3 where instantaneously the following conditions are fulfilled:

$$\begin{aligned} &V_{gsi}(t_1) = V_{gsi}(t_2) = V_{gsi}(t_3) \\ &V_{dsi}(t_1) = V_{dsi}(t_2) = V_{dsi}(t_3) \\ &\dot{V}_{gsi}(t_1) \neq \dot{V}_{gsi}(t_2) \neq \dot{V}_{gsi}(t_3) \\ &\dot{V}_{dsi}(t_1) \neq \dot{V}_{dsi}(t_2) \neq \dot{V}_{dsi}(t_3) \end{aligned} \tag{20}$$

The need for three independent measurements increases significantly the minimum number of large-signal measurements necessary to generate a complete non-linear model. This minimum number can be optimised by exploiting all the degrees of freedom of the measurement set-up, such as DC bias, fundamental frequency, power level, use of multi-tone excitations, etc. [21, 22]. Not only the measurement conditions, but also the required modeling accuracy influences the number of necessary measurements. The reason is that there are several orders of magnitude difference between the elements to be extracted (device currents are typically in the mA range, while device capacitances are typically in the fF range), which implies that the three V_{gsi} and three V_{dsi} time derivatives have to be significantly distinct.

This extraction procedure has been applied to an nMOSFET [13]. The analysis above assumes that the pads and dispersive access transmission lines have been de-embedded from the measurements. Figure 14 presents the extracted I_{dsi} and C_{11} as function of the gate-source voltage. The accuracy of these results can be assessed by comparing them to results obtained by standard DC and *S*-parameter measurement based extractions on the same device. Figure 14 shows a very good agreement, which validates the developed large-signal measurement based non-linear extraction procedure.

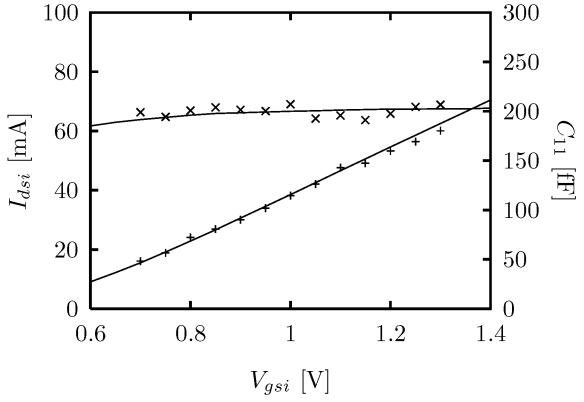


Figure 14. Comparison of S-parameter measurement based (solid line) and large-signal measurement based I_{dsi} (+) and C_{11} (x) of an 0.18 µm × 146 µm nMOSFET at $V_{ds} = 0.8$ V.

4. Conclusions

In this Chapter, several modeling approaches making use of high-frequency measurements have been discussed.

It has been explained that typical approaches for III–V compound devices can be adjusted to MOSFETs. An important change is the more complicated de-embedding of the pads and access transmission lines due to the resistive silicon substrate.

It has also been shown that non-linear models can be determined from either small- or large-signal high-frequency measurements. Concerning the latter, there is a choice between optimization and extraction based approaches, and between equivalent circuit and behavioral model representations. The 'golden' modeling approach does not exist. Depending on the type of device and on the application, one or the other method can be preferable. As a rule of thumb, the following guidelines can be kept in mind:

S-parameter measurement set-ups are common and well spread. If measurement time is not a real issue (i.e., performing an S-parameter sweep across a dense bias grid takes several hours), the classical small-signal measurement based method can be considered. Its advantage is that a non-linear model is obtained that is valid across a wide bias range. Moreover, as both the small- and large-signal equivalent circuit elements are calculated explicitly, feedback to device processing is possible.

A drawback however is that the method requires that S-parameters be measured at all possible biases that can be reached instantaneously during large-signal operation, which is not always possible due to risk for device degradation. In case of large-signal measurements, the experimental operation conditions

can be engineered in such a way that those extreme voltage and current values are only reached instantaneously, which enlarges the potential operation range of the model.

Another surplus value of the large-signal measurements based models is that the number of measurements and the measurement time is usually (significantly) less. This is especially the case for the optimization-based method, which recommended use is to construct quickly a model for a given application. In case a more general model is targeted, the extraction-based approach can be suggested.

As all these models are measurement based, the importance of the measurements should not be underestimated. The design of the experimental conditions at which the measurement data are to be collected is highly important. An example is that the quality of model interpolation is directly related to the grid on which measurements were acquired.

Acknowledgments

The author acknowledges Jan Verspecht and Ewout Vandamme for their contributions to the results as well as for the many interesting discussions. This work has been supported by FWO-Vlaanderen, IWT, and IMEC.

References

[1] Vandamme, E.P.; Schreurs, D.; van Dinther, C.; Badenes, G.; Deferm, L. "Development of an RF large signal MOSFET model based on an equivalent circuit, and comparison with the BSIM3v3 model", *Solid-State Electron.*, **2002**, *46(3)*, 353–360.

[2] Verspecht, J.; Debie, P.; Barel, A.; Martens, L. "Accurate on wafer measurement of phase and amplitude of the spectral components of incident and scattered voltage waves at the signal ports of a nonlinear microwave device", *IEEE MTT-S Int. Microwave Sympos.*, **1995**, 1029–1032.

[3] Schreurs, D.; Verspecht, J. "Large-signal modelling and measuring go hand-in-hand: accurate alternatives to indirect S-parameter methods", *Int. J. RF Microwave Comput.-Aided Engrg.*, **2000**, *10(1)*, 6–18.

[4] Schreurs, D.; Rutkowski, J.; Beyer, A.; Nauwelaers, B. "Development of a frequency-domain simulation tool and non-linear device model from vectorial large-signal measurements", *Int. J. RF Microwave Comput.-Aided Engrg.*, **2000**, *10(1)*, 63–72.

[5] Schreurs, D.; Wood, J.; Tufillaro, N.; Barford, L.; Root, D. "Construction of behavioural models for microwave devices from time-domain large-signal measurements to speed-up high-level design simulations", *Int. J. RF Microwave Comput.-Aided Engrg.*, **2003**, *13(1)*, 54–61.

[6] Jansen, P.; Schreurs, D.; De Raedt, W.; Nauwelaers, B.; Van Rossum, M. "Consistent small-signal and large-signal extraction techniques for heterojunction FETs", *IEEE Trans. Microwave Theory and Techniques*, **1995**, *43(1)*, 87–93.

[7] Schreurs, D.; van Meer, H.; van der Zanden, K.; De Raedt, W.; Nauwelaers, B.; Van de Capelle, A. "Improved HEMT model for low phase noise in InP based MMIC oscillators", *IEEE Trans. Microwave Theory and Techniques*, **1998**, *46(10)*, 1583–1585.

[8] Vandamme, E.P.; Schreurs, D.; van Dinther, C. "Improved three-step de-embedding method to accurately account for the influence of pad parasitics in silicon on-wafer RF test-structures", *IEEE Trans. Electron Dev.*, **2001**, *48(4)*, 737–742.

[9] Dambrine, G.; Cappy, A.; Heliodore, F.; Playez, E. "A new method for determining the FET small-signal equivalent circuit", *IEEE Trans. Microwave Theory and Techniques*, **1998**, *36(7)*, 1151–1159.

[10] Schreurs, D.; Baeyens, Y.; Nauwelaers, B.; De Raedt,W.; Van Hove, M.; Van Rossum, M. "S-parameter measurement based quasi-static large-signal cold HEMT model for resistive mixer design", *Int. J. Microwave and Millimeter-Wave Comput.-Aided Engrg.*, **1996**, *6(4)*, 250–258.

[11] Tsividis, Y. *Operation and modelling of the MOS transistor*, McGraw-Hill; **1999**.

[12] Root, D.E.; Fan, S. "Experimental evaluation of large-signal modeling assumptions based on vector analysis of bias-dependent S-parameter data from MESFETs and HEMTs", *IEEE MTT-S Int. Microwave Sympos.*, **1992**, 255–258.

[13] Augendre, E.; Rooyackers, R.; Caymax, M.; Vandamme, E.P.; De Keersgieter, A.; Perello, C.; Van Dievel, M.; Pochet, S.; Badenes, G. "Elevated source/drain by sacrificial selective epitaxy for high performance deep submicron CMOS: Process window versus complexity", *IEEE Trans. Electron Dev.*, **2000**, *47(7)*, 1484–1491.

[14] Schreurs, D.; Vandamme, E.; Vandenberghe, S.; Carchon, G.; Nauwelaers, B. "Verification of non-linear MOSFET models by intermodulation measurements under loadpull conditions", *Automatic RF Techniques Group Conf. (ARFTG)*, **June 2000**, 58–62.

[15] Schreurs, D.; Vandamme, E.; Vandenberghe, S.; Carchon, G.; Nauwelaers, B. "Applicability of non-linear modelling methods based on vectorial large-signal measurements to MOSFETs", *IEEE MTT-S Int. Microwave Sympos.*, **2000**, 457–460.

[16] Schreurs, D.; Verspecht, J.; Vandenberghe, S.; Vandamme, E. "Straightforward and accurate nonlinear device model parameter estimation method based on vectorial large-signal measurements", *IEEE Trans. Microwave Theory and Techniques*, **2002**, *50(10)*, 2315–2319.

[17] Angelov, I.; Zirath, H.; Rorsman, N. "Validation of a nonlinear transistor model by power spectrum characteristics of HEMT's and MESFET's", *IEEE Trans. Microwave Theory and Techniques*, **1995**, *43(5)*, 1046–1052.

[18] Zhang, Q.J.; Gupta, K.C. *Neural networks for RF and microwave design*, Artech House, **2000**.

[19] Schreurs, D.; Jargon, J.; Remley, K.; DeGroot, D.; Gupta, K.C. "ANN model for HEMTs constructed from large-signal time-domain measurements", *Automatic RF Techniques Group Conf. (ARFTG)*, **June 2002**, 6.

[20] Schreurs, D.; Verspecht, J.; Nauwelaers, B.; Van de Capelle, A.; Van Rossum, M. "Direct extraction of the non-linear model for two-port devices from vectorial non-linear network analyzer measurements", *Eur. Microwave Conf.*, **1997**, 921–926.

[21] Schreurs, D.; Vandenberghe, S.; Wood, J.; Tufillaro, N.; Barford, L.; Root, D.E. "Automatically controlled coverage of the voltage plane of quasi-unilateral devices", *Automatic RF Techniques Group Conf. (ARFTG)*, **May 2001**, 86–90.

[22] Schreurs, D.; Remley, K.A.; Williams, D.F. "A metric for assessing the degree of device nonlinearity and improving experimental design", *IEEE Int. Microwave Sympos.*, **2004**, 795–798.

Chapter 5

EMPIRICAL FET MODELS

Iltcho Angelov
U. Chalmers
E-mail: iltcho.angelov@mc2.chalmers.se

Abstract: This chapter will cover basics of the Empirical FET Models Implementation in CAD tools. First basic experimental characteristics at DC, like I_{ds} and I_{gs} bias dependence will be discussed. Experimental S-parameter, capacitance and high frequency, thermal, power and dispersion characteristics will be shown. They will be linked with the Small and Large Signal Equivalent circuit of the FET. Examples will be given with some basic FET models as they are implemented in CAD tools. It will also be shown how empirical models can be extended to incorporate physical phenomena like thermal effects and dispersion. Finally, models for MOSFET devices will be highlighted.

Key words: FET Modeling; HEMT Modeling.

1. Introduction

The RF performance of FET devices has been dramatically improved in recent years. Today, state of the art FET technology offers very high frequency of operation with high output power. A significant amount of work has been done in the field of high frequency FET transistor modelling and parameter extraction [1–64]. As the output power and operating frequency increase, we face the problem of how to model the high frequency and high power limitations in FET performance and how to implement this in software packages.

Physical modelling approach is very important to optimizing the device structure and to tailor the transistor characteristics for specific application. Nowadays, physical simulators are much faster and more accurate. In the future they will become fast enough to be used in directly for circuit design and

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, 121–155.

better integrated in the microwave designers software tools. When the device is finally available from the processing lab quite often characteristics are different from the simulated. In addition, there are always processing tolerances even when a good and stable process is used. These tolerances can influence the accuracy of all simulations including the accuracy of prediction of the output power, but mainly the accuracy of harmonics and inter-modulation simulations. A problem with physical simulators is that they need detailed data for the material and wafer structure and manufacturing details, which are not always available from the foundries. That is why it is common practice to work with the measured device characteristics. When using experimentally measured device characteristics to extract model, there are two approaches:

2. Equivalent Circuit Approach: Evolution

Direct measurement based approach for modelling FET devices was put on track by D. Root and co-authors [17–20]. Later this approach was refined by number of researchers [56–62]. Nowadays this approach is implemented and used in the software packages. The extracted model is very accurate and provides good description of device characteristics. A problem with this approach is that the model is difficult to extend beyond the regions of measured operating voltages and frequencies. The mounting environment should be kept as in the measurements. When device (or environment) is changed, a complete set of measurements should be done and the model should be extracted again.

Years ago, modelling of semiconductor devices was started using equivalent circuit approach. The explanation is simple- software design tools started from analyzing simple lumped element circuits. When computing power and knowledge were available, it was possible to assemble simple small signal device models in the CAD tools. Figure 1 shows such a simple FET equivalent circuit. The model is a set of lumped passive components – resistors, capacitors and inductances. Their placement and values should correspond to the device physics and geometry parameters of the device. The output current source with

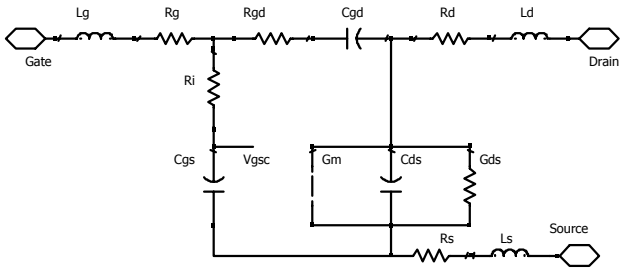


Figure 1. Small Signal Equivalent Circuit of a FET.

transconductance g_m is controlled by the voltage V_{gsc} on the input capacitor C_{gs}. The equivalent circuit approach gives a possibility to extend the model prediction well above the measurements range and when some parameter is changed it is easy to tune the model.

Approximately at the same time several very good works on the small signal FET model and extraction appeared and their extraction procedure to find parameters of the equivalent circuit (EC) is in wide use today [14–16]. This is, because their EC approach is based on the device physics, it is simple and easy to understand and very accurate. For good quality FET, the small signal (SS) model extracted in this way is accurate within 2–5% with the measurements. The extraction is rather simple and when the data are organized in a proper way, the extraction can be done automatically even using directly the software tool:

$$\begin{aligned}\left[Y^i\right] &= \begin{bmatrix} Y^i_{11} & Y^i_{12} \\ Y^i_{21} & Y^i_{22} \end{bmatrix} \\ &= \begin{bmatrix} \dfrac{jC_{gs}\omega}{1+jR_iC_{gs}\omega} + \dfrac{jC_{gd}\omega}{1+jR_{gd}C_{gd}\omega} & -\dfrac{jC_{gd}\omega}{1+jR_{gd}C_{gd}\omega} \\ \dfrac{g_m e^{-j\omega\tau}}{1+jR_iC_{gs}\omega} - \dfrac{jC_{gd}\omega}{1+jR_{gd}C_{gd}\omega} & g_d + jC_{ds}\omega + \dfrac{jC_{gd}\omega}{1+jR_{gd}C_{gd}\omega} \end{bmatrix}\end{aligned} \tag{1}$$

$$\begin{aligned} C_{gs} &= \frac{1}{\omega} \times \mathrm{Im}\left(\frac{1}{\left(Y^i_{11}+Y^i_{12}\right)}\right)^{-1} & C_{gd} &= -\frac{\mathrm{Im}\left(Y^i_{12}\right)}{\omega} \times \left[1+\left(\frac{\mathrm{Re}\left(Y^i_{12}\right)}{\mathrm{Im}\left(Y^i_{12}\right)}\right)^2\right] \\ R_i &= \mathrm{Re}\left(\frac{1}{Y^i_{11}+Y^i_{12}}\right) & R_{gd} &= -\frac{\mathrm{Re}\left(Y^i_{12}\right)}{\mathrm{Im}\left(Y^i_{12}\right)} \times \left[1+\left(\frac{\mathrm{Re}\left(Y^i_{12}\right)}{\mathrm{Im}\left(Y^i_{12}\right)}\right)^2\right]^{-1} \\ g_m &= \left|\left(\frac{Y^i_{21}-Y^i_{12}}{Y^i_{11}+Y^i_{12}}\right)\right| \times \mathrm{Im}\left(\frac{1}{Y^i_{11}+Y^i_{12}}\right)^{-1} & \tau &= -\frac{1}{\omega} \times \left[\arg\left(\frac{Y^i_{21}-Y^i_{12}}{Y^i_{22}+Y^i_{12}}\right) + \frac{\pi}{2}\right] \\ g_d &= \mathrm{Re}\left(Y^i_{22}\right) + \mathrm{Re}\left(Y^i_{12}\right) & C_{ds} &= \frac{1}{\omega} \times \mathrm{Im}\left(Y^i_{22} + Y^i_{12}\right) \end{aligned} \tag{2}$$

When the small signal model and extraction were established and implemented in the CAD tools, the next step was to integrate the small signal equivalent circuit model into the large signal model (LS). Many of the elements of the equivalent circuit are bias dependent and the extended, LS equivalent circuit approach was the simplest way to increase the complexity of the device models. With LS model is possible to include these bias dependencies of nonlinear elements. This provides a possibility to do accurately more complicated tasks like designing nonlinear circuits such as power amplifiers, mixers, oscillators multipliers etc. First, IV characteristics were added to the simulated parameters

and the Small Signal S-parameters were generated directly from the LS equivalent circuit. It is natural to expect that S-parameters generated from the LS FET model with small input power should be equal to the S-parameters generated from the SS equivalent circuit.

3. Current Models

3.1. I_{ds} Current

Extracting the current part of the model is very important part of creating the FET large signal model. Before starting any detailed measurements and modeling it is good to evaluate the quality (functionality) of the selected transistor. It is important to measure or compensate the cable and DC line losses before any extraction starts, especially with currents above 0.1 A. The reason is that, it is impossible to distinguish the influence of external resistances on the IV from the influence of intrinsic device resistances. This problem is common for every kind of device – FET or HBT, that is why, the resistances of the measurement setup should be evaluated carefully before any model extraction is started.

The drain current is measured in wide range of biases sweeping both V_{gs} and V_{ds} as Figure 2. Typically we will need at least 10 gate voltages and 5 to 10 drain voltages depending on the voltage and power range of the transistor. When measurements and extraction are done properly, we can expect that at low frequency, where the contribution from reactive components (capacitance and inductances) is small, the model will be correct. In case low-frequency dispersion phenomena are present in the device, an extended model is required (see Section 6.2). Figure 2a shows typical dependencies of $I_{ds}, G_m\, f\,(V_{gs}, V_{ds})$ for GaAs FET. Figure 2b shows typical g_m dependence vs. V_{gs} for V_{ds} above

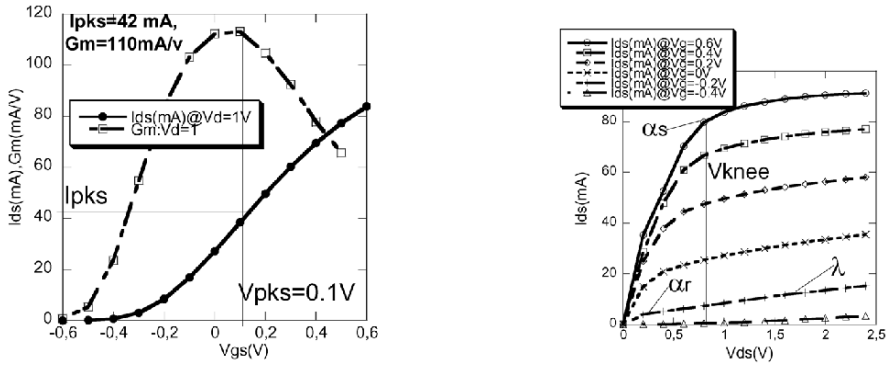


Figure 2. (a) I_{ds}, G_m vs. V_{gs}, (b) I_{ds}, vs. V_{ds}, FET $W = 200\,\mu$m.

knee voltage. The gate voltage V_{pks}, g_m and the drain current I_{pks} at which the maximum transconductance occurs can be used to link measured and modeled I_{ds}. Typically, this inflection point occurs at the gate voltage for which we have the half of the channel current I_{pks}.

For drain voltage above the knee voltage V_{knee} and gate voltage $V_{gs} \cong$ 0.6–0.8 V for GaAs FET the drain current will saturate and reach the maximum channel current. This maximum channel current depends on the material structure, doping profile etc. For GaAs FET the maximum channel current is 0.3–0.5 A/mm and for new material structures like GaN the maximum channel current can be as large as 1.6 A/mm.

When we change the drain voltage, there is a change of the gate voltage for which we have maximum of the transconductance V_{pk} as can be seen on Figure 2. At low drain voltage $V_{ds} = 0.2$ V, the peak of G_m is at $V_{gs} = -0.1$ and at high $V_{ds} > V_{\text{knee}}$ the $V_{pk} = 0.1$ V. Above V_{knee} there is some increase of the drain current, due to the channel opening from the drain voltage influence. If the drain voltage is further increased, breakdown can occur. Typically, high power devices are biased for high efficiency operation i.e., at high voltages and low currents. A properly constructed load line will keep the devices away from the breakdown area and they will be switched from high voltage and low current to high currents and low voltages (close to the V_{knee}). If this is the case, there is no sense to spend much time making very detailed and accurate breakdown model. Only if the device will be operated in the breakdown area it worth spending time to make detailed and accurate breakdown model.

Transconductance and the ratio $P_1 = G_m / I_{ds}$ also change when the drain voltage is changed. This means that the models should have a functional dependence for the peak voltage $V_{pk} = f(V_{ds})$, $P_1 = f(V_{ds})$ to describes the changes of V_{pk}, G_m due to drain voltage influence. Figure 3 shows the I_{ds} vs. V_{gs} dependence when the stepping drain voltage V_{ds} from negative to positive.

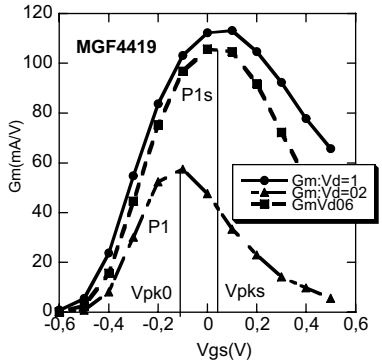


Figure 3. G_m vs. V_{gs} FET 200 μm.

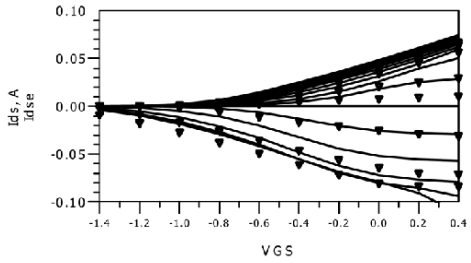


Figure 4. Measured and modeled I_{ds} vs. V_{gs} Symmetrical model.

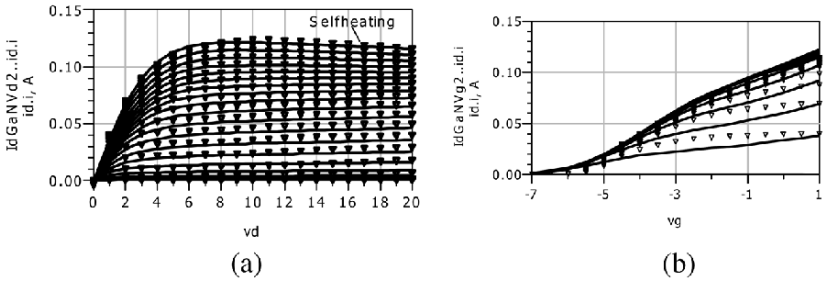


Figure 5. (a) I_{ds} vs. V_{ds} with V_{gs} as a parameter of a GaN FET; (b) I_{ds} vs. V_{gs} with V_{ds} as a parameter of a GaN FET.

As can be seen, the device is not completely symmetrical and this is in part due to the shift of V_{pk} when V_{ds} is negative.

Often due to large device size, highly dissipated power and dispersive effects, the modelled IV characteristics are far from ideal, Figures 4, 5. The self-heating will decrease the drain current at high dissipated power [64]. The decrease of I_{ds} at high dissipated power will critically depend on the thermal resistance R_{therm} and for high power devices it is important to select a proper material with a high thermal conductivity, to make a good thermal design of the transistor – i.e., using properly placed via hols thermal shunts and thin substrate. The technology for the new GaN and SiC devices is very promising, but still not settled and there is substantial activity to improve these devices. We can expect that the IV curves and all parameters for these new, high power devices will gradually become better then for devices with established technology like GaAs.

The basis for the FET operation are two dependencies- the carrier velocity and carrier concentration, Figures 6, 7. Their bias and temperature dependencies will be the main factors which will determine the transistor behavior. The I_{ds} vs. V_{gs} dependence is similar to the carrier concentration dependence vs.

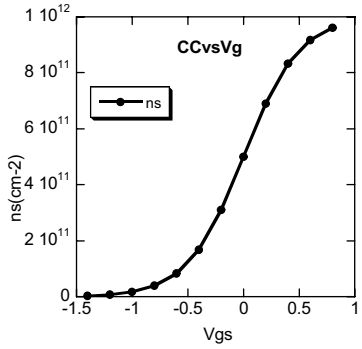


Figure 6. Variation of the 2DEG (ns) sheet densities.

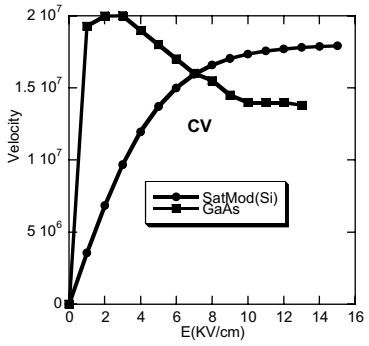


Figure 7. Velocity vs. electric field for GaAs and Si for AlGaAs. GaAs MODFET vs. V_{gs}.

gate voltage, Figure 7 and corresponding modeling function should be selected. Generally, the solution of the Schrödinger and Poisson equation are ***erf*** type of functions, but ***erf*** function is usually not available in circuit simulators. That is why, it can be replaced with other suitable, like ***tanh*** which is accurate enough for this application [7].

In GaAs FET devices at some electric field (V_{ds}, V_{gs}) we observe a maximum of the carrier velocity and transconductance. In Si we have gradual increase of the carrier velocity, which will produce quite different shape of I_{ds}, G_m, G_{ds} as in Figure 8 in comparison with the GaAs Figure 3. The g_m for the Si CMOS device increases with the drain voltage increase and will change shape I_{ds} vs. V_{gs} as well. The different shape of g_m for Si CMOS will produce different harmonic content in comparison with the GaAs FET. This means that in the FET models we should have respective parameters describing these dependences.

There are some general requirements for the selection of the modeling functions in the empirical models. In FET and HBT the device parameters can

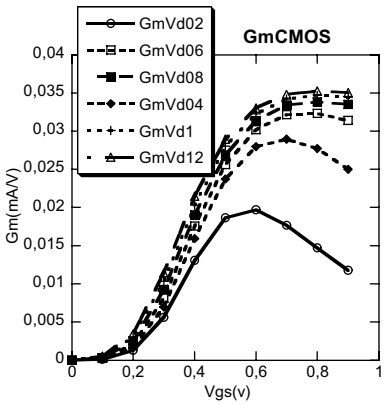


Figure 8. G_m vs. V_{gs} CMOS.

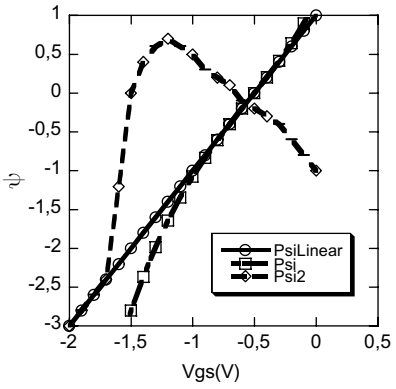


Figure 9. Extracted argument Ψ vs. V_{gs}.

be considered dependent on two voltages $I = f_1(V_{gs}) \cdot f_2(V_{ds})$ or respective V_{be}, V_{ce}. The best solution from extraction and user understanding point of view is to make both parts f_1 and f_2 completely independent – this will greatly simplify extraction. However, when follows from device physics that we have inter-coupling between the $f_1(V_{gs}) \cdot f_2(V_{ds})$ parts, this should be implemented in a proper way. Then, with very small number of additional parameters the model will describe the device behavior accurately. When proposed modeling function is correct and the device is ideal, from the measured data we should obtain a linear function for the extracted argument of f_1 *or* f_2. The derivative will be equal to the measured derivative as in Figure 9. If from the reverse extraction we can get two values of the argument, as this is shown for the *example* function P_{si2} (i.e., we have a $\partial\Psi^2/\partial V_{gs}^2 = 0$) this is an indication that our choice for modeling function is not very good. This is because the selected function

P_{si2} will work in the simulations, but will create problems in the extraction. This is valid also for the sub-functions responsible for the inter-coupling between f_1 and f_2. For example, if the function we guess is $y = Ax^2$ this will work well in the simulation. But obviously there is a problem in the reverse extraction, because the same value of y can be produced by two values of the argument $x = \pm\sqrt{y/A}$.

Often the device is not ideal and we need some flexibility to tune the model. It seems logical that a complex model is more likely to be accurate. This is correct, within limits, because we should always keep in mind that there are processing tolerances and there is no sense making model 1% accurate when process tolerances are 10%. The representation of the Argument as a Power Series (APS) will give a possibility to fit variety of devices. Fitting a polynomial function is rather simple task, but even in this case, parameters of the APS should be selected properly. For example, when we have a negative second term in APS we should always add positive 3-rd term and so on. This will exclude the possibility of a local maximum and dual argument reading and provide required trimming.

3.2. Gate Current

Sometimes we forget that FET devices have gates and ignore that the FET can exhibit significant gate current when driven with high input power. A reason users do not like gate models is that gate current I_{gs} dependence vs. V_{gs} is exponential and this creates problems with the harmonic balance convergence when large number of harmonics is considered. For this reasons the gate current model should be carefully implemented in the software package, properly extracted and used.

In the standard diode equation, $I_{gs} = I_s(\exp(V_{gs}/V_t \cdot N_e) - 1)$, I_s is extracted at $V_{gs} = -\infty$, i.e., at very small currents and very negative V_{gs} for which we do not operate the device. We can change the reference (extracting) point rearranging the diode equation. In the new definition, parameters are taken directly at the typical operating point at high gate current. This can be the knee of I_{gs} vs. V_{gs} characteristics at $V_j = 0.8\,\mathrm{V}$ which is typical GaAs device. The exponent can be limited with some limited function like in Eq. (4b):

$$\begin{aligned} I_{gs} &= I_j(\exp(P_{be}) - \exp(P_{be0})), \\ P_{be} &= P_{be1}((V_{gs} - V_j), P_{be0} = -P_{be1}(V_j), \\ P_{be1} &= q_e/K_b \cdot T_{ambK} \cdot N_{e1} = 1/V_t \cdot N_{e1} \cong 38.695/N_{e1}, \end{aligned} \tag{4a}$$

$$I_{gs} = I_j(\exp(P_{be1}\tanh(V_{gs} - V_j)) - \exp(P_{be1}\tanh(P_{be0}))) \tag{4b}$$

where q_{e^-} is the electron charge, K_{b^-} is the Boltzmann constant, N_{e1} is ideality factor, I_j is measured I_{gs} at V_j [52].

When the transistor is biased as a low noise or small signal amplifier, the gate current is small (well below 1 μA) and can be ignored.

4. Empirical FET Models: Evolution

4.1. Curtice Quadratic Model [11, 52–54]

4.1.1. Standard model

One of the first MESFET model implemented in the software packages was the Curtice FET model [11, 52–54]. The model is very simple, but includes all important transistor parameters – pinch off voltage, transconductance parameter β etc, Eqs. (1)–(5). The model describes well the transconductance and gain with the parameter β, output conductance via parameter λ etc. Due to simplicity and easy to understand and extract, the model is in wide use un general cases, because the model provides a good accuracy predicting gain, output power etc.

$$I_{ds} = \beta(V_{gst} - V_{t0})^2 * \tanh(\alpha * V_{ds}) * (1 + \lambda * V_{ds}); \quad (5)$$
$$\text{for} V_{gsi} \geq 0 \text{ and } I_{ds} = 0 \text{ for } V_{gst} < 0;$$
$$V_{gst} = V_{gsi}(t - T) - (V_{t0} + \gamma \cdot V_{dsi}); \quad (6)$$

Parameter β is transconductance parameter, α define the slope of I_{ds} vs. V_{ds} in the linear region ($V_{ds} < V_{kn}$). λ is the slope in the saturated region ($V_{ds} > V_{kn}$). V_{t0} is the pinch-off voltage. In the CAD tool implementation it is important to set the I_{ds} current equal to 0 for V_{gs} voltages less then pinch-off voltage V_{t0}. There are changes and improvements of the model equations in order to be implemented in the software packages [52–54].

4.1.2. Extended model: Curtice cubic model [52–54]

Later the model was extended with 3-rd term in the polynomial function [52–54] to improve fit for the 3-rd harmonic:

$$I_{ds} = (A_0 + A_1 \cdot V_x^2 + A_2 \cdot V_x^2 + A_3 \cdot V_1^3)\tanh(\gamma * V_{ds}); \quad (7)$$
$$V_1 = V_{gs}(t - \tau)(1 + \beta \cdot (V_{\text{out0}} - V_{ds}));$$
$$\text{for} \quad V_{gsi} - V_{t0} \geq 0 \quad \text{and} \quad I_{ds} = 0 \quad \text{for} \quad V_{gs} - V_{t0} < 0;$$

A_0, A_1, A_2 are polynomial coefficients for the I_{ds} vs. V_{gsi} dependence, V_{out0} is the drain voltage β is extracted.

4.2. Materka-Kacprzak Model [12, 52–54]

A model implemented in simulators soon after the Curtice model was Materka-Kacprzak model [12, 52–54], Eq. (10). The model addresses several important issues – ability to change the transconductance slope with the parameters E_e and K_e and change of the slope of the output conductance with the parameter S_s, K_g.

$$\begin{aligned} I_{ds} = {} & I_{dss}(1 - V_{gsi}(S_s \cdot V_{dsi}/I_{dss}) \\ & \times(1 - V_{gsi} \cdot (t - T)/V_{t0} + \gamma V_{dsi})^{(E_e + K_e \cdot V_{gsi}(t-\tau))} \\ & * \tanh(S_l * V_{dsi}/(I_{dss} \cdot (1 - K_g \cdot V_{gsi}(t - T)), \\ & \text{for} \quad V_{gsi} - V_{t0} \geq 0 \quad \text{and} \quad I_{ds} = 0 \quad \text{for} \quad V_{gst} - V_{t0} < 0; \end{aligned} \tag{10}$$

where I_{dss} is the saturation drain current, V_{t0}– threshold voltage, E_e exponent defining the dependence of saturated current, K_e description of dependence on gate voltage, K_g dependence on V_{gs} of the drain slope in linear region, S_l linear slope of $V_{gs} = 0$ drain characteristic, S_s saturation region drain slope at V_{gs}.

4.3. Triquint Model [21, 52–54]

The major companies like Triquint and Agilent also created FET models and help to extract these models.

In the Triquint model [21] controlling gate voltage is defined as ln (exp (V_{gs})) Eq. (11):

$$\begin{aligned} & I_{ds} = I_{ds0}/(1 + \Delta I_{ds0} V_{dsi}); V_{gst} = V_{gsi}(t - T) - V_{t0} + \gamma^* V_{dsi} \\ & I_{ds0} = (\beta/1 + U V_{gsi}) V_g \cdot K_{\tanh} \\ & V_g = Q V_{st} \cdot \ln(\exp(V_{gst}/Q \cdot V_{st}) + 1); \\ & V_{st} = (N_g + N_d \cdot V_{dsi}) V_t \\ & K_{\tanh} = a \cdot V_{dsi}/(1 + a \cdot V_{dsi}^2)^{0.5} \end{aligned} \tag{11}$$

where I_{ds0}, β is transconductance parameter,V_{t0} pinch-off voltage, U mobility degradation parameter, γ slope of the pinch-off voltage, Q-Power low parameter, N_g Sub-threshold drain parameter, N_d, sub-threshold drain parameter, ΔI_{ds0} Slope of drain characteristics in the saturated region, α slope of drain characteristic un the linear region, T-Channel transit time delay.

4.4. EESOF Model [52]

This is very complete model and is frequently used by foundries, it is supported by complimented extraction programs. Part of model equations is given

by Eq. (12). The model addresses different issues like changing the shape of the transconductance G_m, influence of V_{ds} on G_m and output characteristics etc:

$$\begin{aligned} V_{ts} &= V_{ch} + (V_{ts0} - V_{ch})/(1 + \gamma(V_{ds0} - V_{ds}); \\ I_{ds0} &= G_m \max\left[V_{ch} + V_x(V_{gs}) - ((V_{g0} + V_{t0})/2)\right]; \\ g_{m0} &= G_m \max\left[1 + \gamma(V_{ds0} - V_{ds})\right]; \\ g_{ds0} &= -G_m \max \gamma(V_{gs} - V_{ch}); \end{aligned} \tag{12}$$

4.5. Chalmers FET Model [27, 52–54]

The basic idea in this model is to connect and use directly measured parameters in order to simplify modeling and extraction Eq. (13a). It is supported by complimented extraction programs. The model equations are with continuous derivatives, without poles from $-\infty$ to $+\infty$, without switching or conditioning. The model is optimized to work in the saturation region for $V_{ds} > V_{\text{knee}}$ and V_{gs} for the peak of the transconductance. For saturated V_{ds} and $V_{gs} = V_{pk0}$ the function $\tanh(\alpha V_{ds})(1 + \lambda V_{ds}) \simeq$; $(\lambda \ll 1)$, and the drain current is $I_{ds} = I_{pk}$ by definition. The parameter $P_1 = g_m/I_{pk}$, will automatically define the FET transconductance g_m at this point. Parameters V_{pk0}, I_{pk}, $P_1 = g_m/I_{pk}$ are taken directly from the measurements and as result, the extraction is very simple i.e., 3 parameters $> I_{pk}, V_{pk0}, P_1$ at saturated V_{ds}. The model and derivatives are strictly defined at V_{pk0} and in the vicinity of V_{pk0} where the maximum of the transconductance occurs. For wider range of drain voltages V_{ds} two more parameters $\alpha, \boldsymbol{\lambda}$ are used:

$$\begin{aligned} I_{ds} &= I_{pk}(1 + \tanh(P_{1m}((V_{gs} - V_{pk0}))\tanh(\alpha V_{ds})(1 + \lambda V_{ds}) \\ &\simeq 1; (\lambda \ll 1) \end{aligned} \tag{13a}$$

$$I_{ds} = I_{pk} \quad \text{at} \quad V_{pk0}, G_m = I_{pk} * P_1; \tag{13b}$$

The parameter $\boldsymbol{\alpha}$ together with R_d (and all DC transmission line resistances in the measurement setup) will define the slope of I_{ds} vs. V_{ds} at small drain voltages $V_{ds} < V_{\text{knee}}$. The parameter $\boldsymbol{\lambda}$ will define the slope of I_{ds} vs. V_{ds} at high $V_{ds} > V_{\text{knee}}$ and is extracted at small currents to avoid the influence of the self-heating. These two parameters are common for many models.

For devices with complicated doping profile more sophisticated model structure can be used. The gate dependence is described as a power series using more terms in the power series as P_2, P_3 to track variety of I_{ds} vs. V_{gs} gate dependences. The parameter P_2 will introduce asymmetry of the I_{ds} vs. V_{gs} and will influence the second harmonic and parameter P_3 will trim drain current at gate voltages close to the pinch off and influence the 3-rd harmonic. Typically three terms are enough to provide accuracy better then 5%. As it follows from experimental data, some of parameters like V_{pk}, P_1 are

bias and temperature dependent and in order to have a global model, they are modeled [27], Eq. (15) as:

$$I_{ds} = I_{pk}(1 + \tanh(\Psi_p))\tanh(\alpha V_{ds})(1 + \lambda V_{ds} + \lambda_{sb} \cdot e^{V_{dg}}); \quad (14)$$

$$\psi_p = P_{1m}((V_{gs} - V_{pk0}) + P_2(V_{gs} - V_{pk0})^2 + P_3(V_{gs} - V_{pk0})^3);$$

$$P_{1m} = g_{mpk}/I_{pk};$$

$$V_{pk}(V_{ds}) = V_{pk0} + \Delta V_{pks}\tanh(\alpha_s V_{ds}) - V_{sb2}(V_{dg} - V_{tr})^2; \quad (15)$$

$$\alpha = \alpha_r + \alpha_s[1 + \tanh(\psi_p)]; \quad P_{1m} = P_{1s}(1 + B_1/\cosh(B_2 \cdot V_{ds}));$$

Parameter V_{pk} describes the change of V_{pk} due to the drain voltage, and parameters α_r and α_s change the slope of I_{ds} at small V_{ds}. A good fit in the area of small or negative drain voltages can be important for circuits working at low V_{ds} like resistive mixers, switches etc. The parameters are rather independent in adjusting I_{ds}. For example α_r will influence the drain current at small V_{ds} and small currents, and α_r will influence the drain current at small V_{ds}, and high currents, close to the knee, Figure 2b. Above knee the slope of I_{ds} vs. V_{ds} is adjusted with parameter λ. Breakdown modeling, if required, can be treated with parameters V_{tr}, L_{sb} and V_{sb2} [27, 52–54].

Many of these parameters are typical for all FET. For example, transconductance parameter P_1 for MESFET's is typically $P_1 = 1.2$–1.5, $P_1 = 2 > 4$ for the HEMT, $P_1 = 0.3$ for GaN, $P_1 =$ for 2 for LDMOS etc. High value of P_1 will produce higher gain for the same current, which is good for low noise and high gain applications. But if P_1 is very large, the gate voltage swing (input power) can be limited and this will influence the linearity and inter-modulation characteristics. Transistors with low P_1 like MESFET's, GaAs HEMT's specially designed for linear applications, SiC and GaN FET will have better inter-modulation properties, but lower gain. This means that some compromise should be made if we want to have high efficiency high power and linear amplifier. Depending on the application we can select the best P_1 for our application. Nowadays the physical simulators are fast enough and can help to optimize the device structure for specific application. In Table 1 are given some basic data for different FET devices.

Normally we operate the devices at positive drain voltages and it seems obvious that there is no need to look at negative V_{ds}. When drive level is small this is correct, but when the device is used as power amplifier, switch or mixer, the instantaneous drain voltage is swinging into the negative V_{ds} region. i.e., the drain current model should describe properly the I_{ds} at negative V_{ds} even if the device is biased with positive V_{ds}. Usually, in the circuit simulators the model switching at negative V_{ds} is arranged in a simple way. When the drain voltage V_{ds} is positive the gate voltage V_{gs} controls the drain current. When V_{ds} is negative, the control voltage is switched to V_{gd} and I_{ds} current is calculated from the same equation with reversed sign (I_{ds} is negative). If the device is symmetrical, this is correct. But at the switching point $V_{ds} = 0$ will be a singularity and the

Table 1.

Parameter	MESFET	HEMT	HighGain HEMT	Linear HEMT	SiC	GaN	LDMOS
Ichan [A/mm]	0.3−0.6	0.3−0.6	0.17−0.25	0.3−0.6	0.35	1.5	0.75
P_1	1.1−1.5	2−3	4.5−5.5	1.5	0.1	0.3	2
V_{pk}	−0.5	−0.2	+0.05	−1.4	−9	−3	3.5
V_{knee}	0.75	0.75	0.75	0.75	9	4	3
α_s	1.3−1.5	2−2.5	3.7	1.5	0.14	0.4	1.5
Cap [pF/mm]	1	1	1.0	1.3	0.8	0.7	0.6

derivative of I_{ds} is not defined. As a consequence, it will be more difficult for the HB to converge and the results of the simulations can be wrong in the vicinity of $V_{ds} = 0$. A solution to this is a continuous, single model equation for I_{ds} valid for all control voltages from $-\infty$ to $+\infty$.

For cases like switches and resistive mixers applications, operating at low and negative V_{ds} (as in Figure 4) the drain current equation Eq. (16) is composed from two sources I_{dsp} and I_{dsn}, and which are controlled respectively by V_{gs} and V_{gd} [52]:

$$
\begin{aligned}
I_{ds} &= 0.5(I_{dsp} - I_{dsn}); \qquad (16)\\
I_{dsp} &= I_{pk}(1 + \tanh(\Psi p))(1 + \tanh(\alpha V_{ds})) \cdot (1 + \lambda V_{ds} + \lambda_{sb}.e^{V_{dg}-V_{tr}}),\\
I_{dsn} &= I_{pk}(1 + \tanh(\Psi n))(1 - \tanh(\alpha V_{ds}))(1 - \lambda V_{ds}),\\
\psi_p &= P_{1m}((V_{gs} - V_{pk0}) + P_2(V_{gs} - V_{pk0}) + P_3(V_{gs} - V_{pk0})^3),\\
\psi_n &= P_{1m}((V_{gd} - V_{pk0}) + P_2(V_{gd} - V_{pk0}) + P_3(V_{gd} - V_{pk0})).
\end{aligned}
$$

When V_{ds} is 0 the currents $I_{dsp} = I_{dsn}$ and the drain current $I_{ds} = 0$.

There are cases with when the device has very complicated I_{ds} vs. V_{gs}, V_{ds} dependencies and it is very difficult to obtain a good correspondence between the model and measurements. In this case the power series can be replaced with a data set calculated from measured data [28] i.e. combining both the empirical equivalent circuit models with table based models [17–20] or using the Table Based Model. Using mixed Empirical-Table Approach is possible to combine and extract the best from both. The Empirical Model is serving as envelope for the Table Based Model and the problem with spline function selection, out of the measurement region extension and convergence are solved. This is because, a correct spline functions i.e., FET model equations are used as a spline. The derivatives are continuous and correct and the model will converge well. The linear extrapolation out of the measured data range will be adequate, because the empirical model will limit the solution. The model will be limited

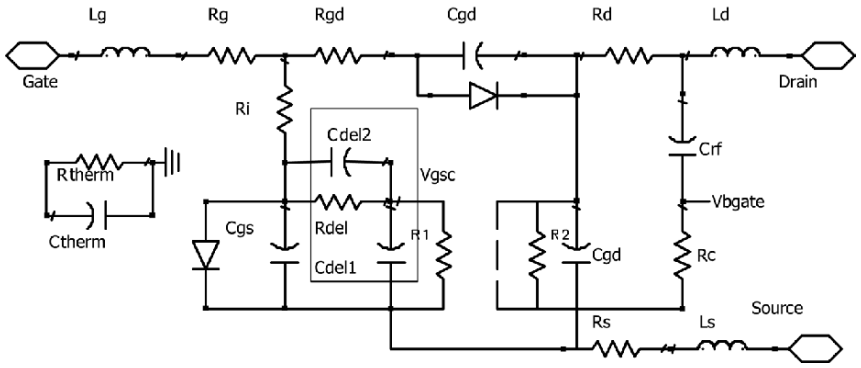


Figure 10. Large Signal Equivalent circuit of the transistor.

and valid out of the measured range, because the data set is naturally limited by using the measured data for the extraction.

Quite often there is spread of parameters and it is important to give the users some flexibility to tune basic model parameters in the Empirical or mixed Empirical – Table Based Model. For example there are always some tolerances in gm, pinch-off voltage, thermal resistance etc. and the model can be arranged in such a way that the user, without making complete measurement and extraction set can change only the required parameter. This can be done with a proper arrangement of the Mixed Empirical – Table Based Model. The Mixed Empirical Table Based Model can be arranged to access the basic parameters $I_{pk}, V_{pk}, P_1, \lambda$, capacitances combining benefits of the Empirical and the Table-Based models. The LS Model is extracted for a typical device, but later it should be possible to trace the process tolerances etc.

The FET large signal equivalent circuit with reactive components included is rather standard, Figure 10. Linear are considered most of the elements and nonlinear (bias dependent) are considered I_{gs}, I_{ds} and capacitances C_{gs} and C_{gd}. The difference between the simple small signal equivalent circuit Figure 1 and LS equivalent circuit are diodes at the gate drain current source, thermal and delay sub-circuit. They are described in more detail in the following sections.

5. Capacitance Models

5.1. Charge Conservation

In multiple extraction and physical simulations on different FET structures was evaluated that the main device capacitances are bias dependent on both voltages $C_{gs} = f(V_{gs}, V_{ds})$ and $C_{gd} = f(V_{gd}, V_{ds})$, Figures 11, 12. This is normal

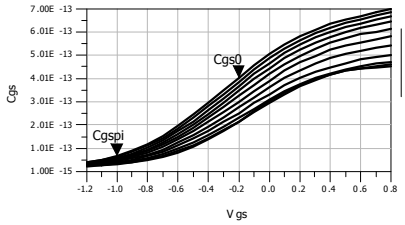


Figure 11. C_{gs} vs. V_{gs}, V_{ds} parameter.

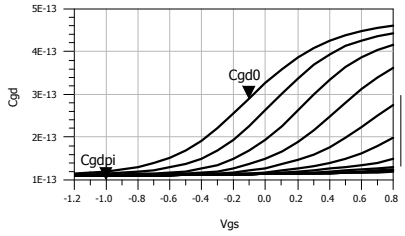


Figure 12. C_{gd} vs. V_{gs}, V_{ds} parameter.

to expect, the problem is how to implement this in the circuit simulators. The charge implementation and conservation problem is very old, several good works are devoted on the topic and propose solutions [4, 45–47]. Traditionally FET total gate charge has been model by two nonlinear charges: gate-source Q_{gs} and gate-drain charge Q_{gd}. A consequence of the dependence of the capacitances on the remote voltage is that we need additional charge control element which D. Root called transcapacitances [17–20].

There are several ways to implement the gate charges into two individual components: Division by capacitances, division by Charge [4].

As FET devices have both gate to source capacitance C_{gs} and gate to drain C_{gd}, it seems natural to use them directly. In this case:

$$C_{gs} = \frac{\partial Q_g}{\partial V_{gs}}; \quad C_{gd} = \frac{\partial Q_g}{\partial V_{gd}} \tag{19}$$

In the case we use capacitances in the implementation, the currents I_s and I_d depend only on the time derivative of their own terminal voltage and not on the changes in any remote voltage. The resulting small signal equivalent circuit is completely consistent with the large signal equivalent circuit and requires no transcapacitances.

Another option is to divide the gate charge Q_g into two independent charges. Then:

$$Q_g = Q_{gs} + Q_{gd} \tag{20}$$

where both Q_{gs} and Q_{gd} are functions of V_{gs} and V_{gd}. Differentiating Q_g with respect to time gives:

$$\begin{aligned} I_g &= I_s + I_d \\ I_s &= \frac{\partial Q_{gs}}{\partial t} = \frac{\partial Q_{gs}}{\partial V_{gs}}\frac{\mathrm{d}V_{gs}}{\mathrm{d}t} + \frac{\partial Q_{gs}}{\partial V_{gd}}\frac{\mathrm{d}V_{gd}}{\mathrm{d}t}; \\ I_d &= \frac{\partial Q_{gd}}{\partial t} = \frac{\partial Q_{gd}}{\partial V_{gs}}\frac{\mathrm{d}V_{gs}}{\mathrm{d}t} + \frac{\partial Q_{gd}}{\partial V_{gd}}\frac{\mathrm{d}V_{gd}}{\mathrm{d}t} \end{aligned} \tag{21}$$

In this case the reactive source and drain currents result from both capacitances and transcapacitances and both definitions charge and capacitance are not equivalent.

A common approach to implement the charge part of every transistor model is to use directly the charge approach. In this case the current of the capacitance is easy to calculate by taking the time derivative of the charge – i.e., multiplying by $j\omega$. This operation is very reliable, because making the derivative will always produce only one solution. This works very well with capacitance which depends only on their own terminal voltage. The problem with all FET transistors is that the gate capacitance depends on the two controlling voltages. When we multiply by $j\omega$ we are making in fact the full derivative of the charge and the end result is not correct if the charge is obtained as integrating the capacitance equation by the terminal voltage. It is obvious that partial (considering the remote part constant) and full derivatives are different. This can be shown with the case of the capacitance model using Eqs. (22–25). Integrating the C_{gs} capacitance by the terminal voltage V_{gs} we obtain Eq. (26). It is assumed that V_{ds} part is constant. If ordinary charge approach is used, multiplying by $j\omega$ will bring obviously different results. i.e. we need to compensate the difference due to the partial derivative – we need an extra term the transcapacitance [4, 17–20, 45–47].

In some advance simulators, for the compiled models, the derivatives of the charges are calculated analytically using the selected terminal voltage. Then the problem is solved in a better way in the sense that the CAD tool is making the derivative vs. respective terminal voltage, considering the remote voltage constant. In this case we will have the capacitance described as a derivative of the charge at the terminal voltage and the capacitances calculated by both methods should be similar.

In the first case we need a correct description of the charge which will compensate for the difference between the partial and full derivative otherwise the model will not be charge conservative. The consequence that the model is not charge conservative is that this difference will create additional current, solution will become path dependent and the HB of the simulator will have difficulties to converge [4, 17–20, 45–47].

5.2. Capacitance Expressions

Figure 11 shows the typical shape of the C_{gs} and C_{gd} capacitances. When the device is symmetrical, for $V_{ds} = 0$ capacitances C_{gs} and C_{gd} are equal. For gate voltage voltages close to pinch off capacitances C_{gs} and C_{gd} have their minimum values C_{gspi} and C_{gdpi} and this should be used in the capacitance models to define the capacitance at the pinch-off. Increasing V_{gs} will increase C_{gs} and C_{gd}. Generally, when V_{ds} increase C_{gs} will increase and saturate at voltages around $V_{ds} = 2\,\mathrm{V}$. In general, the shape of capacitance dependencies will depend on the doping profile and material and in some specific cases a special capacitance model can be developed.

A reasonably good description of the capacitance shape for FET can be obtained using Eqs. (22)–(25) [28, 52–54]:

$$\psi_1 = P_{10} + P_{11} * V_{gs} + P_{111} * V_{ds}; \quad \psi_2 = P_{20} + P_{21} * V_d \tag{22}$$

$$\psi_3 = P_{30} - P_{31} * V_{ds}; \psi_4 = P_{40} + P_{41} * V_{gd} - P_{111} * V_{ds} \tag{23}$$

$$\begin{aligned} C_{gd} = {} & C_{gdp} + C_{gd0} * (1 - P_{111} + \tanh[\psi_3]) \\ & * (1 + \tanh[\psi_4] + 2 * P_{111}) \end{aligned} \tag{24}$$

Independently of the implementation (Capacitance or Charge) and the type of model, in order to have the capacitance model charge conservative it is **mandatory** to fulfil following basic requirement:

$$\frac{\partial C_{gs}}{\partial V_{gd}} = \frac{\partial C_{gd}}{\partial V_{gs}} \tag{25}$$

This means that the equations for the capacitances C_{gs} and C_{gd} should be symmetrical and model coefficients should be selected properly. In the case of Eqs. (22)–(24) this means that $P_{11} = P_{41}$ and $P_{22} = P_{33}$. The consequences can be non-convergence in the HB. A good test for the consistency of the capacitance models is to simulate the S-parameters in the small signal case and S-parameters simulated in the LS case with HB, but with very small input power. If this difference is small, this means that the capacitance model is correct and implemented properly. For capacitances described with Eqs. (19), (20) the charges are:

$$\begin{aligned} Q_{gs} = {} & \int C_{gs} * \partial V_{gs} = C_{gsp} * V_{gs} + C_{gs0} * (\Psi_1 + Lc1 - Q_{gs0}) \\ & * (1 + \tanh[\Psi_2]))/P_{11} \\ Lc1 = {} & \log[\cosh(\psi_1)]; \quad Lc10 = \log[\cosh(P_{10} + P_{111} * V_{ds})] \\ Q_{gs0} = {} & P_{10} + P_{111} * V_{ds} + Lc10 \end{aligned} \tag{26}$$

$$
\begin{aligned}
Q_{gd} &= \int C_{gd} * \partial V_{gd} = C_{gdp} * V_{gd} + C_{gd0} \\
&\quad * (\Psi_4 + Lc4 - Q_{gd0}) * (1 - P_{111} + \tanh[\Psi_3]))/P_{41} \\
Lc4 &= \log[\cosh(\psi_4)]; \quad Lc40 = \log[\cosh(P_{40} + P_{111} * V_{ds})] \\
Q_{gd0} &= P_{40} + P_{111} * V_{ds} + Lc40
\end{aligned}
\tag{27}
$$

The functions for capacitances, charges and their derivatives are symmetrical and defined from $-\infty < V_{gs}, V_{gd}, V_{ds} < +\infty$. A problem that should be accounted is the boundary condition problem. – i.e., what will be with the capacitances (charges) when the capacitance terminal is shorted and there is a voltage on the remote terminal as in Figures 13, 14. For example, when the gate source junction is shorted ($V_{gs} = 0$) the capacitance C_{gs} will continue to exist and the charge Q_{gs} should be $Q_{gs} = 0$ independent from remote voltage V_{ds}. This puts additional constraints on the boundary conditions for the charge definition. For these reasons some circuit simulators use separate Q_{gs}, Q_{gd}, but taking into account the boundary condition with charges Q_{gs0} and Q_{gd0}. As it can be seen from Figures 13, 14, when $V_{gs} = 0$ the charge $Q_{gs} = 0$ and when $V_{gd} = 0$ the charge $Q_{gd} = 0$ independently from the remote voltage V_{ds}.

Generally the most circuit simulators use either standard charge approach or direct capacitance approach.

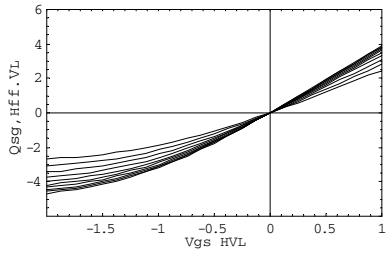


Figure 13. Charge Q_{gs} vs. V_{gs}.

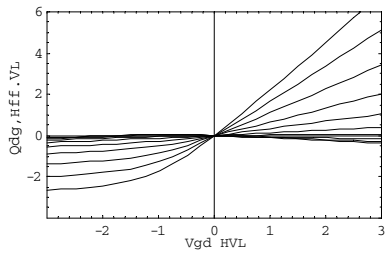


Figure 14. Charge Q_{gd} vs. V_{gd}, V_{ds} parameter.

It is important to know that **always should be some small difference** in the calculated S(Y)-parameters depending on the implementation type- capacitance or charge, even if the same model parameters for the capacitances are used. The origin of this difference in the calculated S-parameters depending on the implementations is very well described by S. Maas [4]. As a consequence, it is important to keep the same tape of the model in extraction and later in the circuit simulations, because this small difference can be accounted fitting the S-parameters with the selected capacitance model and fulfilling necessary condition Eq. (25c).

Possible solution to the problem is to use a single gate charge Q_g definition. The total gate charge Q_g is function of V_{gs} and $V_{gd}(V_{ds})$ [28, 49]. When some of these voltages changes, Q_g change as well-the gate current is $\mathrm{d}Q_g/\mathrm{dt}$. In this case, the total gate charge $Q_g = Q_{gs} + Q_{gd}$ and I_g composed by derivatives of the two charges Q_{gs} and Q_{gd}. It follows from this that $I_g = I_s + I_d$.

Where

$$\begin{aligned} I_s &= \frac{\mathrm{d}Q_{gs}}{\mathrm{d}t} = \frac{\partial Q_{gs}}{\partial V_{gs}}\frac{\mathrm{d}V_{gs}}{\mathrm{d}t} + \frac{\partial Q_{gs}}{\partial V_{gd}}\frac{\mathrm{d}V_{gd}}{\mathrm{d}t} \\ I_d &= \frac{\mathrm{d}Q_{gd}}{\mathrm{d}t} = \frac{\partial Q_{gd}}{\partial V_{gs}}\frac{\mathrm{d}V_{gs}}{\mathrm{d}t} + \frac{\partial Q_{gd}}{\partial V_{gd}}\frac{\mathrm{d}V_{gd}}{\mathrm{d}t} \end{aligned} \tag{28}$$

This will work well and the only problem is that we cannot extract charges directly and we need to derive them via capacitances and S-parameters. Because of these complications with the charge definitions and difficulties with implementation in the CAD tools, many circuit simulators use capacitance formulation. As explained, when capacitance approach is used the resulting small-signal equivalent circuit consists of the small signal capacitances evaluated at the corresponding DC voltage.

The first step in the Cap implementation is to calculate the time derivatives $\mathrm{d}V_{gs}/\mathrm{d}t$ and $\mathrm{d}V_{gd}/\mathrm{d}t$ of the respective terminal voltage. i.e. the simulator should calculate the time derivative in reliable way. When the CAD tool is able to make the transient analysis (as most modern CAD tools do), the capacitance type of implementation can be done reliably. The respective current is obtained by multiplying the time derivative with the capacitance equation:

$$I_{gsc} = C_{gs} * \frac{\partial V_{gs}}{\partial t}; \quad I_{gdc} = C_{gd} * \frac{\partial V_{gd}}{\partial t} \tag{29}$$

It is important to arrange the DC component of the time derivative to be equal to 0 within the accuracy of the HB simulations (typ. less then $I_{dc} < 10^{-15}$ A). If implemented in a proper way, this will result in consistent small- and large-signal models and we don't need any trans-capacitances. This because, the time derivatives depend only on their terminal voltage. A problem that can arise using this approach is the convergence in the HB simulations. This can happened, in the first step of calculating the time derivatives if the functions for

the C_{gs}, C_{gd} are not continuous with well-defined derivatives. Using smooth functions with infinite numbers of derivatives without singularities from $-\infty$ to $+\infty$ helps to solve the problem. Another important moment is to implement these operations Eq. (29) in a proper way.

Generally, the convergence problems are caused by poor numerical conditioning of the Jacobian matrix, caused by a combination of very large and very small numerical values. In nearly all new circuit simulators the Krylov solvers are much less robust, when dealing with ill-conditioned matrices, than some of the older solvers without Krylov solvers. So, in the past, some of these things were not a problem, but suddenly now they are.

In the capacitance implementation, problems can be caused by poor numerical conditioning of the Jacobian matrix, due to a combination of very large and very small numerical values.

For example, in the FET model with capacitance formulation we need to generate $\mathrm{d}V/\mathrm{d}T$ and $C(V)$. The derivative $\mathrm{d}V/\mathrm{d}T$ is very large, but $C(V)$ is very small, and when these are put in the Jacobian, the $\mathrm{d}V/\mathrm{d}T$ entries are much larger than other entries, so the matrix solution is poor.

The simplest solution proposed by S. Maas [4] and implemented in Microwave office, AWR is to multiply $\mathrm{d}V/\mathrm{d}T$ by a small number (for example 1e–9) before passing it to the capacitance expression. Then, C(V) is multiplied by the inverse of that number (1e9 in this case). It seems simple, but it will make a lot of difference. It is a good idea to arrange this scaled factor to be accessed in easy way by the user, because the best performance depends on the circuit (derivatives of the charge) and the user can find what is best for his application.

If this is done properly, the FET model with the capacitance implementation can converge better, specially if we keep the DC current via capacitance $I_{\mathrm{cap}} = 0$ in the HB simulations.

6. Recent Extensions

6.1. Thermal Effects

It is known that solid-state devices are temperature sensitive. There two main reasons for the change of the transistor parameters vs. the temperature. The first is the change of carrier concentration vs. the temperature and the second-change of mobility. Both are reduced when the temperature is increased. The reduction of the carrier concentration will reduce the channel current and reduced mobility will produce smaller transconductance at higher temperature for the FET devices, i.e. negative T_{cIpk}, T_{cP1}. The change of the mobility will also influence the speed of the device and in turn change (increase) the device capacitances (positive $T_c C_{gs0}$). This effect is beneficial when the device is used as a small

signal, low noise amplifier – cooling the amplifier will drastically improve the gain and noise performance of the FET amplifier. This is due to increased g_m (gain) and reduced channel noise which are strongly dependent on the channel temperature. The thermal effects are very negative for high power FET devices. The result is significant reduction of the drain current and gain at high operating temperatures and when dissipated power is high. In addition to the effects directly observed (reduction of the current and the transconductance) the RF and dispersion characteristics are also influenced. This is due to the increased influence of the traps at higher temperature. To account for the temperature changes the equations for the currents and charges should be extended with the terms describing the temperature dependencies vs. junction temperature $T_j = R_{\text{therm}} \cdot P_d + T_{\text{amb}}$ where P_d is dissipated power T_{amb} is the ambient temperature. The thermal resistance is generally nonlinear but for simplicity can be considered constant In this case the temperature increase can be modelled as a thermo-electrical circuit consisting of the thermal resistance R_{therm} and the thermal capacitance C_{therm}. The thermal capacitance models the thermal storage of the structure and the thermal constant is $R_{\text{therm}} * C_{\text{therm}}$. When thermal equivalent circuit is used, $T_j = V_{\text{therm}}$ can be treated like any other control voltage and can be found interactively in the HB simulations. i.e., $T_j = T_{\text{amb}} + V_{\text{therm}}$; $P_d = P_{dc} + P_{rf}$. Because the dissipated power contain the RF power P_{rf} the junction temperature will be time dependent. The thermal mass of the chip will filter out the RF temperature variations, but it will not filter the low frequency modulation signal and we can experience so called memory effects.

To account for the basic effects of self-heating we need to make temperature dependent at least several parameters like: I_{pk}, which are connected with the channel current (approximately $I_{\text{chan}}/2$), transconductance *connected* with mobility (parameter $P_1 = g_m/I_{pk}$), and device junction capacitances C_{gs0} and C_{gd0}. In addition to these parameters, for high power devices the delay parameters R_{del}, C_{del} and breakdown parameters should be considered temperature dependent.

If low frequency modulation of the signal is to be considered, dispersion parameters can be made temperature dependent. The temperature dependencies of all these parameters are rather linear in the temperature range $\pm 100^\circ$ C and temperature coefficients are very small. Typically for GaAs FET $T_c I_{pk}$ and $T_{cP1} = -0.025$. Because of this, they can be modeled as linear functions:

$$K = K_0(1 + T_{CK}(T_j - T_{\text{ref}})) \tag{30}$$

where $K = I_{pk}, P_1, C_{gs0}$ and C_{gd0}. T_{CK} is the temperature coefficient of parameter K. The temperature T_j is determined from the total dissipated power and the thermal resistance.

The change of device parasitic resistances is very small vs. temperature and it is usually considered that the resistors temperature should be equal to the device operating temperature.

6.2. Dispersion Modelling

Years ago when first FET were made, the researchers were unsatisfied to find that transconductance gm and output resistance (conductance) R_{ds} are quite different at high frequency in comparison to the DC values. Figure 15 show typical shape of the g_m and g_{ds} vs. frequency. It should be noticed that the effect is concentrated at rather low frequency, typically below 1 kHz and all the changes are usually settled at frequency 5–10 MHz. The interesting thing is that in some HEMT devices is possible find even a small increase of the extracted g_m vs. frequency.

It was found that the reasons for these effects are basically the material and surface defects which are always present. As long as material and device surface have some defects – we will always have dispersive effects.

From the first glance these changes look rather small and seem that they can be ignored. This is correct in some cases, but when the device is working as an oscillator, RF switch, RF modulated high power amplifier these small changes in the output conductance and transconductance will produce significant effects. The oscillator will become noisy, the slope of the switched RF power will be changed and in high power amplifiers memory effects will be visible – i.e., the output will depend in some way on the modulating signal. As usually, these effects are becoming more critical at high temperatures – i.e., will be more critical for high power and high temperature of operation.

Devices which can deliver high power should have high operating current and high breakdown voltage i.e., rather large device size. Due to this, the dispersive effects become more significant, because they are directly proportional to the surface area [29-41]. Dispersive effects will become more significant for devices with new material systems like GaN, SiC, but even for GaAs these effects can be significant. For this reason, a proper implementation of more accurate dispersion models in circuit simulators is becoming important. An additional effect of highly dissipated power is that as the device is operating at

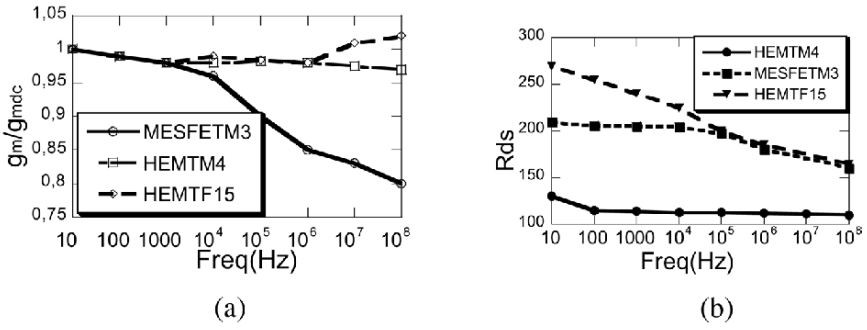


Figure 15. (a) Transconductance gm and (b) R_{ds} vs. frequency.

higher junction temperatures the thermal problems will become more severe, because power is dissipated in a comparably small volume that can be locally overheated. Finally, for large devices the intrinsic delay can cause additional problems. Due to all these effects, at high frequency the high power devices do not deliver the power their DC and small signal S-parameters predict. This can be seen when comparing the maximum tuned output power at different operating frequencies. It is known that this decrease of the maximum tuned power is not due only to the higher losses in the matching circuit and higher resistive losses in the transistor, but largely to the more pronounced physical effects as listed above.

On the topic of correct modeling of the g_m and R_{ds} dispersion are devoted many papers [29-41] and this issue is probably even more important with the new devices like CMOS, GaN. The best is to use an EC based on the physical approach as [29] or back gate approach [30–33], but usually in circuit simulators the simple EC approach is used [34], as shown in Figure 10. In this case a simple R, C branch is used to model the R_{ds} dispersion. The R_c should be bias dependent; otherwise the simulator will not produce correct results for I_{ds}, and Power Added Efficiency at RF. The network with constant R_c will give additional RF current $I_{rf} = V_{ds}/R_c$ and this will produce an extra DC current in the simulations. A correction to the problem can be made making R_c bias dependent and this is the simplest solution implemented in CAD tools:

$$R_{c\min} + R_{c\max}/(1 + \tanh[\psi]) \tag{31}$$

Quite often we forget that the device is symmetrical and dispersion effects existing on the drain side (G_{ds}) exist on the gate side (g_m). Using a similar network at the input R_{cin}, C_{rfin} we can model g_m dispersion, as shown Figure 10.

The best is to organize the model structure in such a way that four terminals are available. The fourth terminal can be used to account for dispersion using the back-gate approach. [30–33]. It is known that this will produce a proper SS description of the g_m and g_{ds} dispersion. If implemented in a proper way in the LS model, this approach works well in both the LS and SS case. This can be done by injecting the feedback RF signal V_{bgate}, shown in Figure 10, directly into the I_{ds} equations, Eq. (15b). From the parasitic coupling, the output RF voltage via C_{rf} and R_c, the backgate voltage V_{bgate} is fed to the gate and controls the drain current at RF. Using this approach, the parameters R_c and C_{rf} will have values close to values we can expect from the device physics.

The modified current equation including the backgate part is [63, 64]:

$$V_{pk}(V_{ds}) = V_{pks} - \Delta V_{pks} + \Delta V_{pks} * \tanh(\alpha_s V_{ds} + K_{BG} * V_{\text{bgate}}); \tag{15b}$$

$$P_{1m} = P_1 * [(1 + \Delta P_1)(1 + \tanh(\alpha_s V_{ds}))]; \tag{32a}$$

$$P_{2m} = P_2 * [(1 + \Delta P_2)(1 + \tanh(\alpha_s V_{ds}))]; \tag{32b}$$

$$P_{3m} = P_3 * [(1 + \Delta P3)(1 + \tanh(\alpha_s V_{ds}))]; \tag{32c}$$

where ψ_p is a power series function centered at V_{pk}. A new term K_{bg} is introduced which controls the intrinsic gate voltage at RF. As it was mention parameters, like V_{pk} and $P_1, P_2, P_3, \ldots$ exhibit bias dependence and this has been accounted by Eq. (32) for the general use.

For high voltage devices, or when very accurate fit for I_{ds} and the harmonics is important, the equations Eq. (32) can provide improved fit, like was already demonstrated in Figure 5 [63]. This is because Eq. (32) gives the possibility to handle both positive and negative changes of the harmonic content. The basic parameters are determined directly from measurements and secondary parameters like P_{2m}, P_{3m}, K_{bg} are optimized with the CAD tool. Such modeling approach allows to use a simple extraction procedure and extracted parameters are trimmed using the CAD tool optimizers.

When dissipated power is small (less then 200 mW) then all the measurements can be done in one sequence, sweeping V_{gs} and stepping V_{ds} and measuring the currents and S-parameters. It is rather important to start measurements from low frequency in order to track the dispersion effects and to improve the accuracy of modelling of the current source.

For high power devices, multiple bias S-parameter measurements should be performed splitting the measurements in two voltage ranges $\rightarrow V_{ds} < V_{\text{knee}}$ and high currents and $V_{ds} V_{\text{knee}} - 30\,\text{V}$ and small currents as in the example in Figure 16. This is needed, because the high power devices operating in class B, C, D, E, F, are usually biased at high voltage and small current, but during the voltage swing they reach very high currents for V_{ds} around the knee voltage. That is why, it is important to evaluate the device along the typical load line. Such a detailed S-parameter evaluation will also provide information on whether the capacitances and their models are behaving properly, because most of the capacitance changes are below and around the knee voltage.

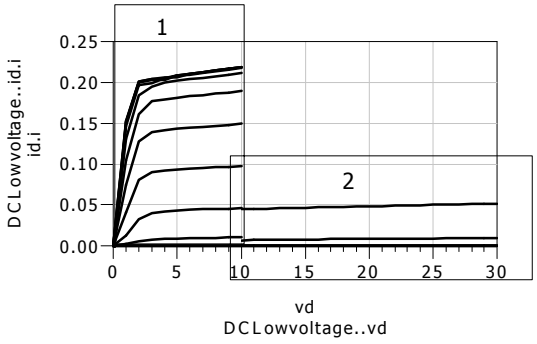


Figure 16. Dual region DC measurements: Region1 High I_{ds}, Low V_{ds}, Region2 Low I_{ds}, High V_{ds}.

6.3. Model Evaluation

It is commonly considered that performing a DC and S-parameter measurements is enough to extract a good quality transistor model. If the goal is to have a model which will predict the gain S-parameters and output power this is correct. Pulsed IV and S-parameter measurements can provide additional info, especially for high power or dispersive devices, but even these data is not enough. If we want to have a model which will predict properly harmonics, then some kind of LS measurements evaluating the harmonic content should be used to trim the model. Only in this case we can be confident that the model will describe the harmonics properly, because the DC and S-parameter evaluation is not enough. We can make very simple simulation experiment with the current source. Usually we are satisfied when the modelling accuracy for the current is better then 5%. We start with a model parameter $P_1 = 2$, $P_2 = 0$, $P_3 = 1.5$. If we change the parameter P_3 which is responsible for I_{ds} characteristics close to the pinch-off and influencing the 3 harmonic to $P_3 = 0.5$, we will see very small change – only 3–4% in the drain current. The same small change DI_{ds} will produce nearly 15 dB difference in the simulated 3-rd harmonic Figure 17. These results are common for every model and every transistor that is why it is important to evaluate the ability of models to describe harmonics with additional measurements.

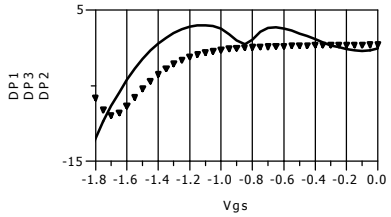


Figure 17. Change of the harmonic output.

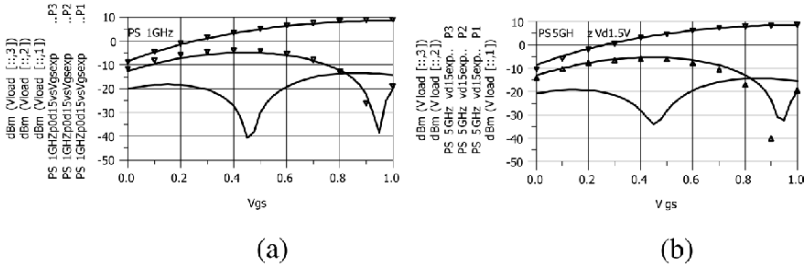


Figure 18. (a) PS measurement results 1 GHz (b) PS measurement results 5 GHz for CMOS device.

The simplest way to evaluate the harmonic contents generated from the device is the direct way to measure harmonics. It is good to evaluate the device at 2 fundamental frequencies – one low frequency – 0.1–1 Ghz depending on the device size to evaluate the nonlinearity of the current source and at high frequency close to the frequency we will operate the device. The masurements should be made sweeping V_{gs} and having as a parameter V_{ds}. Quite often we see that the people are showing P_{out} and harmonics vs input power. It can be shown that nearly every model can be adjusted to give reaonable correspondence, but later they will be surprised to see that the model is not describing harmonics accurately. Typically we need 10 measurements of V_{gs} and several V_{ds}. Figure 18 show some typical results.

6.4. Delay Modelling

The initial hope of researchers that a better model of the dispersion would solve the problem and provide an accurate prediction of the output power at high frequency for high power devices turned out to be false. It was found that even the good fit for the S-parameters does not provide the proper prediction of the output power at high frequency, i.e., it is not able to predict the significant drop of the tuned output power vs. frequency.

By using Large Signal Network Analyzer (LSNA) measurements [63] is possible to observe that the waveforms at high frequency are not efficient any more. The LSNA data provide very important information about the generated waveforms at the tuned condition directly at the device terminal. The model is supposed to reproduce accurately these waveforms.

At low frequency 2 GHz, the waveforms are quite normal, as shown in Figure 19 and Figure 20, and the device delivers 26 dBm at 10 dBm input power. At high frequency, the device is not able to swing to the DC values of the currents, refer to Figure 20b, this phenomenon is called current slump.

For example, at 18 GHz the minimum drain voltage that can be reached at 10 dBm power is 6.3 v, see Figure 20b, in comparison with 0.8 V at 1 GHz, and

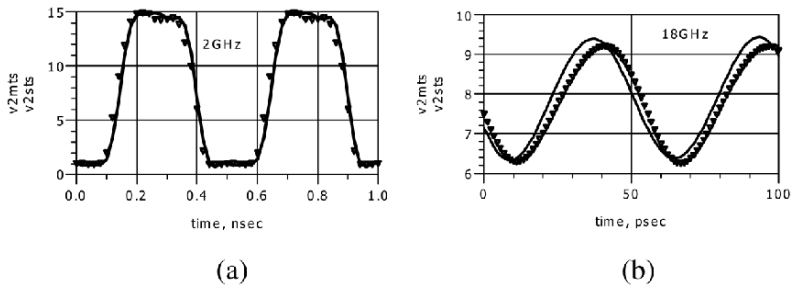
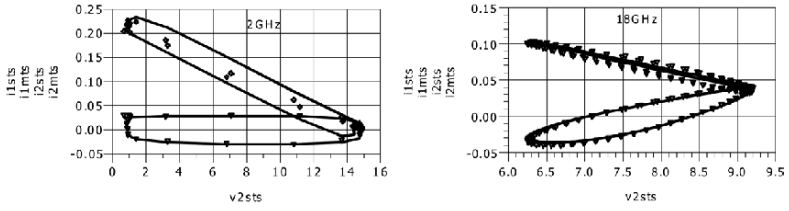


Figure 19. Time waveforms: (a) 2 GHz, (b) 18 GHz.



(a) (b)

Figure 20. Time waveforms I_{gs} (i1), I_{ds} (i2) vs. V_{ds} (v2) (a) 2 GHz, $V_{\text{dmin}} = 0.8$ V (b) 18 GHz $V_{\text{dmin}} = 6.3$ V.

see Figure 20a. The fit for the input current is good, which is a sign that the capacitance are not responsible for this and the capacitance models for C_{gs}, C_{gd} are correct. i.e. the capacitances are not responsible for the loss of power in tuned condition at high frequency.

It can be determined that the voltage V_{gsc} controlling the output current I_{ds} is reduced and delayed thus causing the output waveforms to not be able to follow the input. This was found to be one of the reasons for the low output power (respective low efficiency) at high frequency for high power FET devices.

It is known that in HB simulators is assumed that the model is quasi-static, nonlinear devices are evaluated in time domain and time (frequency) dependent equations for the currents will not behave properly [4, 17]. This means that time-delayed response, explicit frequency dependences of current equations should be avoided. From device physics, the only elements we can use to model the intrinsic part of the devices in circuit simulators are capacitances, resistances and equations connecting the currents and charges. Inductances and layout parameters can be associated with extrinsic part of the device and de-embedded.

In addition, the frequency dependence of the maximum output power is rather complicated and a simple RC network will not provide an adequate fit. After some trials it was found that a delay network (elements C_{del1}, C_{del2}, R_{del}), connected at the input (see Figure 10) provides a good description of these effects [63, 64]. At high frequency, the capacitor C_{del1} shunts the input and directly decreases the magnitude of the control voltage V_{gsc} and introduces the observed delay. The value of the delay capacitance was found by fitting the S-parameters and turned out to be very low, in the order of 2–3 fF. This is so low, that it can be the capacitance of the gate footprint. A possible reason for the delay resistance can be the charging resistance between the 2 Deg. layers and the buffer. The time constant $C_{\text{del}} - R_{\text{del1}}$ will determine the frequency at which the high frequency and high power limitations start to work. The frequency dependence of the output power can be fine tuned using the capacitance C_{del2}. Both delay capacitors C_{del1} and C_{del2} are quite similar, that is why, for simplicity they can be considered equal. The delay network is shunting the

input capacitance C_{gs}, but the values of $C_{\text{del1}}, C_{\text{del2}}$ are so small that they do not significantly influence the input. This means that the ordinary methods to extract bias dependencies of capacitances C_{gs} and C_{gd} can be used.

Thus, the LS model with a back-gate dispersion model and delay and gate control network will work well for small dissipated power and will describe the frequency dependence of the tuned maximum power and large signal gain accurately. Even a simple linear temperature-dependent model for R_c, R_{del} and C_{del} improves the fit, but a better fit can be obtained if more complicated thermal resistance model is arranged from 2 thermal resistors R_{therm1} and R_{therm2} connected in series. In this case R_{therm1} will describe the overheating occurring in a narrow volume, and R_{therm2} will describe the thermal resistance between the volume in which the power is generated and the heat sink.

The output capacitance C_{ds} will critically influence the output power at high frequency. That is why the reduction of all parasitic capacitances is important if the goal is to create a broadband high power amplifier.

7. Empirical CMOS Model

Similar approach can be used to model CMOS devices, taking into account the specific effects for the CMOS device. For example, the I_{ds} current close to pinch-off gate voltages (i.e., very small currents) is very close to exponential as can be seen from logarithmic plot Figure 21. This means that a corresponding term should be available in the current equation Eqs. (28),(29).

The CMOS devices are inherently symmetric and this means that the symmetric I_{ds} model should be used, but modified for CMOS [55]. If it is very important to have a very good accuracy at small V_{ds}, then it is recommended to use V_{ds} bias dependent P_2 and P_3 as in Eq. (32).

Usually for RF application is not required very high accuracy at small V_{ds} and small currents. If this is important, then the special attention should be paid for the fit at small currents, using the parameter for the exponent λ_1. The number of parameters for I_{ds} is low and most of them can be determined directly from

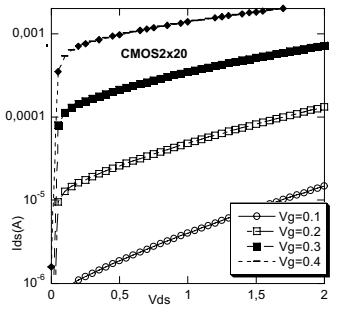


Figure 21. I_{ds} vs. V_{ds} - small current V_{gs} bias.

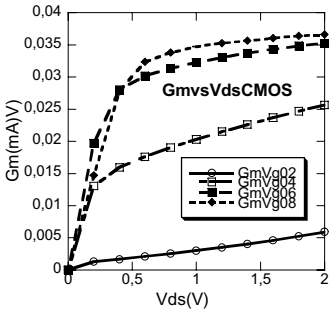


Figure 22. G_m vs. V_{ds} CMOS.

measurements and the remaining parameters are extracted using optimization default CAD tool optimizers.

$$I_{ds} = 0.5(I_{dsp} - I_{dsn})\ldots \tag{12}$$

$$I_{dsp} = I_{pk}(1 + \tanh(\psi_p))(1 + \tanh(\alpha_p V_{ds})) \times (1 + \lambda_p V_{ds} + \lambda_{1p}\exp(((V_{ds}/V_{kn}) - 1))) \tag{33}$$

$$I_{dsn} = I_{pk}(1 + \tanh(\psi_n))(1 + \tanh(\alpha_n V_{ds})) \times (1 - \lambda_n V_{ds} - \lambda_{1n}\exp(((V_{ds}/V_{kn}) - 1))) \tag{34}$$

where $\psi_{p,n}$ are power series functions centered at V_{pk}.

Typically three terms of the power series are enough to produce I_{ds} model accuracy of 2–5%. In a similar way V_{pk} and I_{pk} are the gate voltage and the drain current at which the maximum of the trans-conductance occurs, α_r, α_s are the saturation parameters, and the parameter λ accounts for channel length modulation. Drain voltage dependence of parameters, like V_{pk} and λ is described by Eqs. (15), (33).

The equivalent circuit of the CMOS transistor is much more complicated in comparison with ordinary FET, due to the influence of the bulk. In the small signal EQ Circuit there are multiple parasitic coupling pairs $C_{\text{gbulk}}, R_{\text{gbulk}}, R_{\text{sbulk}}, C_{\text{sbulk}}, R_{\text{dbulk}}$, and C_{dbulk} [55]. These parasitic couplings will affect the FET behavior mainly at RF frequency. The bulk influence at DC and low RF is handled using the backgate approach with parameter K_{bg} in the equation for V_{pk}.

The CMOS capacitances are different from the MESFET and HEMT capacitances. For this reasons the CMOS capacitance model was proposed which track closer the measured dependencies [55], Eqs. (35)–(36):

$$C_{gs} = C_{gsp} + C_{gs0}(1 + V_{gs} + P_{10})/((P_{11} + (V_{gs} - P_{10})^2))^{0.5})(1 + \tanh[P_{20} + P_{21}V_{ds}]) \tag{35}$$

$$C_{gd} = C_{gdp} + C_{gd0}(1 + V_{gd} + P_{40})/((P_{41} + (V_{gd} - P_{40})^2))^{0.5})(1 + \tanh[P_{30} - P_{31}V_{ds}]), \tag{36}$$

The selected functions for C_{gs}, C_{gd} are symmetric with well-defined derivatives. This results in good fit in the S-parameters, and very good convergence behaviour in HB.

Acknowledgments

The author is grateful to colleagues from Chalmers, AWR, Mitsubishi, Agilent, Ansoft for their help and constant support in the modeling work, as well as for the outstanding discussions and very positive feedback. A special thanks goes to H. Zirath, N. Rorsman, E. Kollberg, M. Fernadhl, C. Fager, K. Andersson, S. Maas, A. Inoue, S. Goto, K.Choumei, T. Hirayama, D. Root, D. Schreurs, J. Verspecht.

Table of abbreviations

CMOS	complementary MOS
MOS(FET)	metal oxide semiconductor (field effect transistor)
RF	radio frequency
DC	direct current
CAD	computer aided design
MMIC	Monolithic microwave integrated circuit
LSNA	Large signal network analyzer
HB	Harmonic Balance
I-V	current-voltage

Table of symbols

V_{gs}, V_{ds}, V_{gd}	gate-to-source voltage, drain-to-source voltage, gate-to-drain voltage
I_{ds}	drain-to-source current
V_{gsi}, V_{dsi}	intrinsic gate-to-source and drain-to-source voltages
g_m, g_{ds}	transconductance, output conductance
C_{gs}, C_{gd}, C_{ds}	gate-to-source, gate-to-drain, and drain-to-source capacitances
C_{pd}, C_{pg}	access capacitances at the drain and gate, resp.
R_g	gate resistance
R_i, R_{gd}	resistances for the NQS modeling
R_g, R_s, R_d	resistances at the gate, source and drain, resp.
L_g, L_s, L_d	inductances at the gate, source and drain, resp.
Q_g	total gate charge

Q_{gs}	gate-source charge
Q_{gd}	gate-drain charge
$I_{pk}, V_{pk}, \mathrm{Pi}, \alpha_i, \lambda_i,$	drain current fitting parameters

References

[1] Liou, J.J.; Schwierz, F. "RF MOSFET: recent advances and future trends" *Electron Dev. and Solid-State Circuits, 2003 IEEE Conf.*, **December 16–18, 2003**, 185–192.

[2] Schwierz, F.; Liou, J.J. "Development of RF transistors: a historical prospect solid-state and integrated-circuit technology, 2001". *Proceedings 6th International Conference on Electron Devices Volume 2*, **October 22–25, 2001**, *23*, 1314–1319.

[3] Lopez, J.M. *et al*. "Design optimization of AlInAs-GaInAs HEMTs for high frequency applications!", *IEEE Trans. Electron Dev.*, **April 2004**, *51(4)*, 521–528.

[4] Maas, S. Nonlinear Microwave and RF Circuits, Artech House, **2003**.

[5] Anholt, R. "Electrical and thermal characterization of MESFETs, HEMTs, and HBTs", Artech House, **1995**.

[6] Nguyen, L.D.; Larson, L.; Mishra, U. "Ultra-high-speed MODFET: A tutorial review", *Procs. IEEE*, **1992**, *80(4)*, 494–499.

[7] Rohdin, H.; Roblin, P. "A MODFET DC model with improved pinch off and saturation characteristics", *IEEE Trans. Electron Dev.*, **1986**, *33(5)*, 664–672.

[8] Johnoson, R.; Johnsohn, B.; Bjad, A. "A unified physical DC and AC MESFET model for circuit simulation and device modeling", *IEEE Trans. Electron Dev.*, **1987**, *34(9)*, 1965–1971.

[9] Weiss, M.; Pavlidis, D. "The influence of device physical parameters on HEMT large-signal characteristics", *IEEE Trans. Microwave Theory Tech.*, **1988**, *36(2)*, 239–244.

[10] Rauscher, C.; Willing, H.A. "Simulation of nonlinear microwave FET performance using a quasi-static model", *IEEE Trans. Microwave Theory Tech.*, **October 1979**, *27(10)*, 834–840.

[11] Curtice, W. "A MESFET model for use in the design of GaAs integrated circuit", *IEEE Trans. Microwave Theory Tech.*, **1980**, *28(5)*, 448–455.

[12] Materka, A.; Kacprzak, T. "Computer calculation of large-signal GaAs FET amplifiers characteristics", *IEEE Trans. Microwave Theory Tech.*, **1985**, *33(2)*, 129–135.

[13] Brazil, T. "A universal large-signal equivalent circuit model for the GaAs MESFET", *Proc. 21st Eur. Microwave Conf.*, **1991**, 921–926.

[14] Dambrine, G.; Cappy, A. "A new method for Determining the FET small-signal equivalent circuit", *IEEE Trans. Microwave Theory Tech.*, **July 1988**, *36(7)*, 1151–1159.

[15] Berroth, M.; Bosch, R. "High-frequency equivalent circuit of GaAs FETs for large-signal applications", *IEEE Trans. Microwave Theory Tech.*, **February 1991**, *39(2)*, 224–229.

[16] Berroth, M.; Bosch; R. "Broad-band determination of the FET small-signal equivalent circuit", *IEEE Trans. Microwave Theory Tech.*, **July 1990**, *38(7)*, 891–895.

[17] Root, D.; Hughes, B. "Principles of nonlinear active device modeling for circuit simulation", *#2 Automatic Radio Frequency Technique Group Conf.*, **December 1988**.

[18] Root, D.; Fan; S.; Meyer, J. "Technology-independent large-signal FET models: A measurement-based approach to active device modeling", *15th ARMMS Conf.*, **September 1991**.

[19] Root, D.E. "Measurement-based mathematical active device modeling for high frequency circuit simulation", *IEICE Trans. Electron*, **June 1999**, *E82-C(6)*, 924–936.

[20] Root, D.E. "Nonlinear charge modeling for FET large-signal simulation and its importance for IP3 and ACPR in communication", *Proc. 44th IEEE 2001 Midwest Sympos. Circ. Syst. (MWSCAS)*, **August 2001**, *2*, 768–772.

[21] Hallgren, R. B.; Litzenberg, P. H. "TOM3 capacitance model: Linking large- and small-signal MESFET models in SPICE", *IEEE Trans. Microwave Theory Tech.*, **May 1999**, *47(5)*, 556–562.

[22] Trew, R. J. "MESFET models for microwave CAD applications", *Microwave Millimeter-Wave CAE*, **April 1991**, *1(2)*, 143–158.

[23] Teyssier, J.P.; Viaud, L.P.; Quere, R. "A new nonlinear I(V) model for FET devices including breakdown effects", *IEEE Microwave Guided Wave Lett.*, **April 1994**, *4(4)*, 104–107.

[24] Angelov, I.; Zirath, H.; Rorsman, N. "A new empirical model for HEMT and MESFET devices", *IEEE Trans. Microwave Theory Tech.*, **1992**, *40(12)*, 2258–2266.

[25] Bandler, J.; Zhang, Q.; Ye, S.; Chen, S. "Efficient large-signal FET parameter extraction using harmonics", *IEEE Trans. Microwave Theory Tech.*, **December 1989**, *37(12)*, 2099–2108.

[26] Angelov, I.; Zirath, H.; Rorsman, N. "Validation of a nonlinear HEMT model by power spectrum characteristics", *IEEE MTT-S Digest*, **1994**, 1571–1574.

[27] Angelov, I.; Bengtsson, L.; Garcia, M. "Extensions of the chalmers nonlinear HEMT and MESFET model", *IEEE Trans. Microwave Theory Tech.*, **October 1996**, *46(11)*, 1664–1674.

[28] Angelov, I.; Rorsman, N.; Stenarson, J.; Garcia, M.; Zirath, H. "An empirical table based FET model", *IEEE Trans. Microwave Theory Tech.*, **December 1999**, *47(12)*, 2350–2357.

[29] Kunihiro, K.; Ohno, Y. "A large-signal equivalent circuit model for substrate-induced drain-lag phenomena in HJFET's", *IEEE Trans. Electron Dev.*, **1996**, *43(9)*, 1336–1342.

[30] Conger, J.; Peczalski, A.; Shur, M. "Modeling frequency dependence of GaAs MESFET characteristics", *IEEE J. Solid State Circ.*, **1994**, *29(1)*, 71–76.

[31] Scheinberg, N.; Bayruns, R.; Goyal, R. "A low-frequency GaAs MESFET circuit model", *IEEE J. Solid-State Circ.*, **April 1988**, *23(2)*, 605–608.

[32] Canfield, P.C.; Lam, S.C.F.; Allst, D.J. "Modelling of frequency and temperature effects in GaAs MESFETs", *IEEE J. Solid-State Circ.*, **February 1990**, *25(1)*, 299–306.

[33] M. Lee Forbes, L , "A Self-back-gating GaAs MESFET model for low-frequency anomalies", *IEEE Trans. Electron Dev.*, **October 1990**, *37(10)*, 2148–2157.

[34] Camacho-Penalosa, C.; Aitchison, C. "Modeling frequency dependence of output impedance of a microwave MESFET at low frequencies", *Electron. Lett.*, **June 1985**, *21(12)*, 528–529.

[35] Reynoso-Hernandez, J.; Graffeuil, J. "Output conductance frequency dispersion and low-frequency noise in HEMT's and MESFET's", *IEEE Trans. Microwave Theory Tech.*, **September 1989**, *37(9)*, 1478–1481.

[36] Ladbrooke, P.; Blight, S. "Low-field low-frequency dispersion of transconductance in GaAs MESFETs with implication for other rate-dependent anomalies", *IEEE Trans. Electron Dev.*, **March 1988**, *35(3)*, 257–263.

[37] Kompa, G. "Modeling of dispersive microwave FET devices using a quasi-static approach", *Int. J. Microwave Millimeter-Wave Comput.-Aided Engg.*, **1995**, *5(3)*, 173–194.

[38] Paggi, M.; Williams, P.; Borrego, J. "Nonlinear GaAs MESFET modeling using pulsed gate measurements", *IEEE Trans. Microwave Theory Tech.*, **December 1988**, *36(12)*, 1593–1597.

[39] Teyssier, J.P.; Campovecchio, M.; Sommet, C.; Portilla, J.; Quere, R. "A Pulsed S-parameter measurement set-up for the nonlinear characterization of FETs and bipolar transistors", *Proc. 23rd Eur. Microwave Conf.*, **1993**, 489–493.

[40] Curtice, W.R.; Bennett, J.R.; Suda, D.; Syrett, B.A. "Modelling of current lag in GaAs IC's", *IEEE MTT-S Int. Microwave Sympos. Digest*, **June 1998**, *2*, 603–606.

[41] Anholt, R.; Swirhun, S. "Experimental investigation of the temperature dependence of GaAs FET equivalent circuits", *IEEE Trans. Electron Dev.*, **September 1992**, *39(9)*, 2029–2036.

[42] Fukui, H. "Thermal resistance of GaAs FET", *Proc. IEDM*, **1980**, 118–121.

[43] Lee, K.; Shur, M. "A new interpretation of "End" resistance Measurements", *IEEE Electron Dev. Letters*, **January 1984**, *5(1)*, 5–6.

[44] Debie, P.; Martens, L. "Fast and accurate extraction of parasitic resistances for nonlinear gas MESFET device models", *IEEE Trans. Electron Dev.*, **December 1995**, *42(12)*, 2239–2242.

[45] Snider, A.D. "Charge conservation and the transcapacitance: An exposition", *IEEE Trans. Edu.*, **November 1995**, *38(4)*, 376–379.

[46] Calvo, M.; Snider, A.; Dunleavy, L. "Resolving capacitor discrepanses between large and small signal models", *IEEE Trans. Microwave Theory Tech.*, **June 1995**, 1251–1254.

[47] Kalio, "A new rule for MESFET gate charge division", *Int J. Circ. Theory. Appl.*, **2004**, *32*, 139–165.

[48] Cojocaru, V.I.; Brazil, T.J.; "A scalable general-purpose model for microwave FETs including DC/AC dispersion effects", *IEEE Trans. Microwave Theory Tech.*, **December 1997**, *45(12, part 2)*, 2248–2255.

[49] Wren, M.; Brazil, T.J. "Enhanced prediction of pHEMT nonlinear distortion using a novel charge conservative model", *IEEE MTT-S Microwave Sympos. Digest*, **June 2004**, *1*, 31–34.

[50] Wood, J.; Root, D.E. "A symmetric and thermally de-embedded nonlinear FET model for wireless and microwave applications", *IEEE MTT-S Microwave Sympos. Digest*, **June 2004**, *1*, 35–38.

[51] Osorio, R.; Berroth, M.; Marsetz, W.; Verweyen, L.; Demmler, M.; Massler, H.; Neumann, M.; Schlechtweg, M. "Analytical charge conservative large signal model for MODFETs validated up to MM-wave range", *IEEE MTT-S Microwave Sympos. Digest*, **June 1998**, *2*, 595–598.

[52] ADS User manual, Agilent.

[53] Microwave Office User manual, AWR.

[54] Microwave Designer User manual, Ansoft.

[55] Angelov, I.; Fernhdal, M.; Ingvarson, F.; Zirath, H.; Vickes, H.O. "CMOS large signal model for CAD", *IEEE MTT-S Microwave Sympos. Digest*, **June 2003**, *2*, 643–646.

[56] Filicori, F.; Vannini, G.; Monaco, V.A. "A nonlinear integral model of electron devices for HB circuit analysis", *IEEE Trans. Microwave Theory Tech.*, **July 1992**, *40(7)*, 1456–1465.

[57] Filicori, F.; Mambrioni, A.; Monaco, V.A. "Large-signal narrow band quasi-black-box modelling of microwave transistors", *IEEE Trans. Microwave Theory Tech.*, **June 1986**, *86(1)*, 393–396.

[58] Florian, C.; Filicori, F.; Mirri, D.; Brazil, T.; Wren, M. "CAD identification and validation of a non-linear dynamic model for performance analysis of large-signal amplifiers", *IEEE MTT-S Microwave Sympos. Digest*, **June 2003**, *3*, 2125–2128.

[59] Filicori, F.; Vannini, G.; Santarelli, A.; Mediavilla, A.; Tazon, A.; Newport, Y. "Empirical modeling of low-frequency dispersive effects due to traps and thermal phenomena in III-V FETs", *IEEE MTT-S Microwave Sympos. Digest*, **1995**, *3*, 1557–1560.

[60] Filicori, F.; Monaco, V.A.; Vannini, G. "A harmonic-balance-oriented modeling approach for microwave electron devices", *Electron Dev. Meeting*, **December 1991**, 345–348.

[61] Ghione, G.; Naldi, C.U.; Filicori, F. "Physical modeling of GaAs MESFETs in an integrated CAD environment: From device technology to microwave circuit performance", *IEEE Trans. Microwave Theory Tech.*, **March 1989**, *37(3)*, 457–468.

[62] Santarelli, A.; Filicori, F.; Vannini, G.; Rinaldi, P. "'Backgating' model including self-heating for low-frequency dispersive effects in III-V FETs", *Electron. Lett.*, **October 1998**, *34(20)*, 1974–1976.

[63] Angelov, I.; Inoue, A.; Hirayama, T.; Schreurs, D.; Verspecht, J. "On the modelling of high frequency and high power limitations of FETs", *INMMIC*, **November 2004**, Rome.

[64] Angelov, I.; Desmaris, V.; Dynefors, K.; Nilsson, P.Å.; Rorsman, N.; Zirath, H. "On the large-signal modelling of AlGaN/GaN HEMTs and SiC MESFETs", *Eur. Microwave Conf.*, **2005**, 379–383.

Chapter 6

MODELING THE SOI MOSFET NONLINEARITIES

An empirical approach

B. Parvais[1]*, A. Siligaris[2]
[1]*Université catholique de Louvain (UCL), Microwave Laboratory – Place du Levant, 3; B-1348 Louvain-la-Neuve – Belgium*
[2]*IEMN, ANODE group, CNRS, UMR 8520, Avenue Poincaré, BP 60069, Villeneuve d'Ascq 59652, France*
E-mail: alexandre.siligaris@iemn.univ-lille1.fr

Abstract: In the radio frequency field of applications, the knowledge of both the linear and nonlinear behavior of circuits and devices is required. This chapter is intended to provide simple and efficient models suited for predicting the Silicon-on-Insulator (SOI) MOSFETs nonlinearities for high frequency applications. A specific attention is paid to the floating body effects and a semi-empirical approach is used for its rapid extraction capability. A Volterra based model is first introduced to explain how the MOSFET distortion varies with the frequency. In a second step, a more complete model suited for computer-aided design is presented. Experiments are used to validate the model and to show the linearity of devices presenting or not floating body effects.

Key words: silicon-on-insulator; floating body effects; linearity; distortion; HD; IMD; Volterra; empirical RF modeling; MOSFET modeling.

*B. Parvais is now with IMEC, SPDT/Mixed-Signal Technology Integration group, Kapeldreef 5, B-3001 Leuven, Belgium – email: Bertrand.Parvais@imec.be

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, 157–180.

1. Introduction

In order to follow Moore's law, aggressive scaling has been followed for years, generating high gate leakage and short channel effects [1]. In that context, silicon-on-insulator (SOI) technology appears as an interesting alternative to standard planar bulk devices [2, 3]. Indeed, the SOI structure not only kills the latch-up and improves digital soft-error immunity, but also allows a better control of the channel, leading to an improved sub-threshold slope and lower short-channel effects. In particular, the SOI multiple-gate devices consist of a very promising solution for the fabrication of high performance devices for low-power applications [4].

The buried oxide permits to reduce the parasitic junction capacitances between drain (source) and body, allowing higher operating frequencies. Moreover, the crosstalk immunity is improved [5] and a low level of dielectric losses is achievable – which is a milestone for radio-frequency (RF) applications – when high resistivity SOI substrates are used [6].

The downscaling of MOS technology led to the integration of analog and digital functions on a single chip (SoC). Nowadays, CMOS is considered as the appropriate low cost technology for low GHz telecommunication. In these applications, the transistor nonlinearity characterization and modeling is crucial for circuit design, since it can either be a limiting parameter (e.g., distortion in amplifiers), or a need for good functionality (in mixers for instance).

This chapter is intended to provide SOI MOSFET modeling techniques that are able to describe these nonlinear properties. In order to model the nonlinear behavior of the SOI device from DC to RF coherently, it is necessary to account for the dispersive character of some physical phenomena, such as the floating body (FB) effects. These SOI particularities are introduced in Section 2. In Section 3, a simple and comprehensive analytical model is provided to analyze the evolution of the MOSFET distortion with the frequency. A more general and more complete model dedicated to computer-aided design (CAD) is described in Section 4. Even though several physically based compact models exist [7–9], an empirical approach is proposed in this chapter to permit a quick model extraction, which is desirable when the technology evolves rapidly. The analytical model is introduced to get a comprehensive understanding of the weak nonlinearities under various frequency ranges, while the slightly more complex CAD model is shown to be very efficient in commercial CAD tools for MMICs simulations under both small and large signal analysis.

2. SOI MOSFET Devices

The basic characteristic of SOI technology is the separation of the top active region from the underlying mechanical substrate by a thick insulator layer.

Several techniques exist for the fabrication of SOI substrates (Smart Cut®, SIMOX®, BESOI . . .), however today the Smart-Cut is from far the most commonly used [10].

2.1. Partially- and Fully-Depleted Devices

Depending on the thickness of the top active silicon layer, the film may be fully depleted (FD) or partially depleted (PD) from the majority carriers. In PD devices, the depletion zones of the front gate and the back gate do not interact and a neutral zone exists. This zone is called *body*. When the Si thickness is small enough, the two depletion zones are connected and the device is fully depleted from majority carriers. The Figures 1a and 1b sketch the crosssections respectively of a partially depleted and a fully depleted transistor.

A PD device basically behaves as a bulk device with a floating body. The SOI structure exhibits some advantages (e.g., improved RF capabilities due to smaller parasitic capacitances), but special non-ideal behaviors related to the floating body are experienced. In order to control the body potential, particular structures can be fabricated: for n-channel devices, a lateral p^{++} implant may be used to form an ohmic contact to the transistor body. These implants are generally connected to the source *via* the first two metal layers. In common source configuration, the body potential is thus forced to zero. A schematic top view of the layout of a PD device with the lateral p^{++} implants is showed in

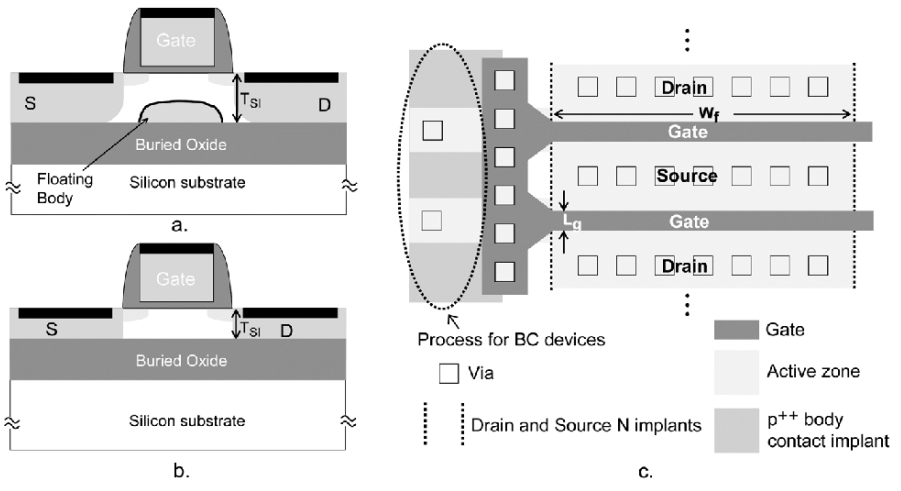


Figure 1. *a.* Schematic lateral cross section of a partially depleted SOI MOSFET. *b.* Schematic lateral cross section of a fully depleted SOI MOSFET. *c.* Schematic top view of multi-finger gate RF n-MOSFET with lateral p^{++} body contacts (BC).

Figure 1c. There are thus two categories in which PD transistors are classified: floating body (FB) and body tied (BT) devices.

The behavior of FD devices (Figure 1b) is closer to the one of an ideal MOS transistor since the gate better controls the channel. Indeed, the coupling between the gate potential and the potential of the channel at the Si-SiO_2 interface is much higher for a FD than for a PD transistor [10]. This results from the difference between the gate capacitance of FD transistors (composed of the series connection of the front gate oxide capacitance), and of PD devices (composed of only the front gate oxide capacitance and the depletion capacitance in series).

2.2. Floating Body Effects

The SOI circuits suffer from several dynamic FB effects as hysteresis and history effects, due to the finite time constant of the generation/recombination mechanisms involved. Two typical physical phenomena are known to induce FB effects; avalanche current (kink effect) and gate leakage (gate-induced floating body effect).

Figure 2a shows the measured DC drain current of two PD devices (FB and BT) having the same technological process and geometrical parameters. We notice a kink in the output characteristic of the FB PD device. This typical floating-body effect can be explained as follows [10].

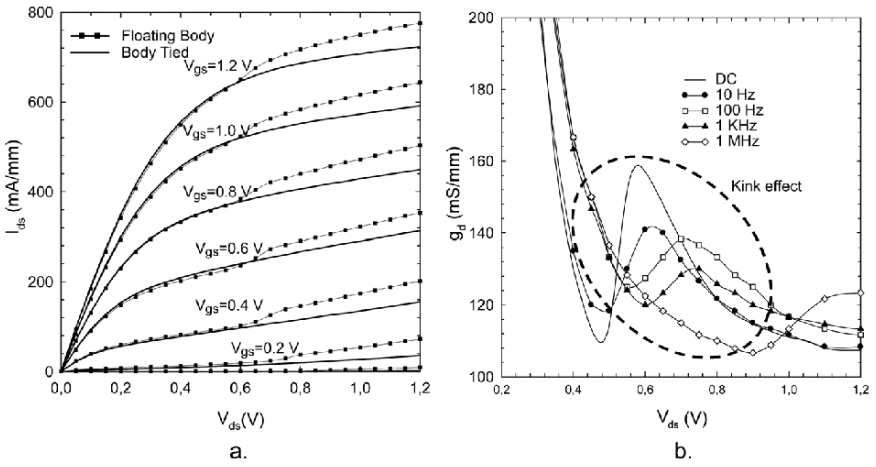


Figure 2. *a.* Measured drain current *versus* the drain voltage for various gate bias values. Symbols: FB. Solid line: BT ($60 \times 1\,\mu m/0.12\,\mu m$). *b.* Measured drain conductance of the FB device *versus* the drain voltage for various applied frequencies. $V_{gs} = 0.6\,V$.

When the drain voltage of a thick-film PD nMOSFET is high enough, electrons can acquire sufficient energy in the high electric field zone near the drain and create electron-hole pairs via the impact ionization mechanism. The generated electrons rapidly move towards the drain, while the holes (which are the majority carriers in the p-type body) migrate towards lower potential i.e., the floating body. The injection of holes into the floating body forward biases the source-body diode. The body potential increases, decreasing the threshold voltage. This in turn induces an increase of the drain current.

It is well known that this kink is frequency-dependent. In Figure 2b, the measured output conductance of a FB device is shown for various applied frequencies. The kink decreases as frequency arises and tends to disappear over about 1 MHz. The frequency dependence relies on AC body voltage filtering through source/body capacitance [11]. This FB effect is known to induce a Lorentzian like low frequency noise overshoot [12].

The kink effect may be attenuated when using intrinsic doping of the channel [13]. Such a low doping concentration is favorable for low-voltage and high-speed applications because it allows to obtain low threshold voltage and slightly higher mobility.

The kink effect can be (almost) eliminated by two means. On one hand, the BT structure may be used, as shown in Figure 2a. Indeed, an ideal body contact can remove the excess majority carriers in the body. In practice, the finite neutral region underneath the gate results in a resistive discharging path, which establishes a potential drop when the excess carriers flow to the lateral body contact. For a finite body resistance, a small kink in the output conductance is thus still observed, but it presents a dispersive behavior at higher frequency [14]. Indeed, the output characteristic is dominated by the resistive body discharge at low frequency, while at higher frequency, the capacitive path combines with this resistance.

On the other hand, the kink effect is not observed in thin-film FD devices at room temperature. Because the full depletion of the film, the source-to-body diode is forward biased, and the holes can readily recombine in the source without raising the body potential.

Finally, gate tunneling intervenes in floating body effects [15, 16]. Indeed, gate tunneling not only increases the device leakage, but leads to charging and discharging the PD body region. To understand this, let us consider a n-type device. Holes are injected into the body through gate tunneling. The body voltage is then increased, resulting to a kink in the transconducance. The AC analysis of this effect showed that it could be described as a function of the frequency by a pole-zero doublet [17]. The pole frequency corresponds to the product of all the resistances and the capacitances seen by the body towards the external nodes. According to experimental data, the zero appears at higher frequency than the pole.

In summary, several mechanisms are particular to the FB structure of SOI devices. These mechanisms have in common a low-frequency dispersion of the transistor output characteristics. While the gate induced FB effect is visible in the triode region, the avalanche induced kink effect occurs in saturation, i.e., the operation mode used in RF applications. Therefore, and for the sake of clarity through this book, only the kink effect will be pointed out. Nevertheless, the same approach can be followed to model the other FB phenomena. The reader interested in the nonlinear behavior of SOI devices in triode regime is reported elsewhere [18].

3. MOSFET Nonlinearities: Figures of Merit and Analytical Model

In this section, a simplified MOSFET analytical model is presented in order to get a comprehensive understanding of the SOI transistor nonlinear mechanisms. It also allows circuit designers to deal with the different trade-offs involved in their design.

3.1. Motivation

The origin of the MOSFET nonlinearities is explained by the semiconductor physics. The DC drain current exhibits a highly nonlinear characteristic when the drain and gate voltages (V_{gs} and V_{ds}, respectively) are varied. In new technologies dedicated to high frequency, the dimensions shrink and other linearity degradations linked to short channel effects appear [19]. So carrier velocity saturation, channel length modulation and mobility degradation have to be carefully described for an accurate large signal modeling of the MOSFET [20] (see Chapter 2). At high frequency of operation, the transistor behavior still depends on the DC current characteristics, but in addition, the reactances affect the behavior of the device.

At low frequency, a Taylor series analysis is generally used to get analytical expressions of the distortion [21], while the Volterra series are used at high frequency when inductors and capacitors play an important role [22, 23, 19]. In these cases, the nonlinear elements are described in terms of the Taylor series expansion of their current-voltage or charge-voltage characteristics. This limits the validity range of these techniques, as the derivatives of the characteristic around any bias point must remain constant over the AC voltage and current swing from that bias point. To ensure the validity of the Taylor series, the nonlinearity must be weak enough and the excitation signals small enough. We speak about *weak* nonlinear systems. This corresponds roughly to the range of input power for which there is no compression. This means in practice that the

input voltage amplitude should typically not exceed 0.3 V to ensure the weak nonlinear behavior of the transistor.

Analytical modeling is on the other hand widely used for its predictive capability. Indeed, while the measurement of the DC current-voltage (I–V) characteristics is today common, performing nonlinear measurements at microwave is a complex task. Instrumentation for nonlinear systems must allow simultaneous multiple tone measurements and requires a calibration of all separate incident and reflected waves instead of single frequency measurements and a relative calibration for linear system. Specific instrumentation setups and experiment design are required. Measurements results involving spectrum analyzer, sampling oscilloscopes and vectorial network analyzers are found in the literature. The recent development of large-signal network analyzers (LSNA) permits an accurate characterization of devices and systems at microwaves [24] (see Chapter 4), at the price of an expensive and non-widespread setup.

In the following of this section, a comprehensive explanation of the MOSFET distortion modeling is provided at both low and high frequencies.

3.2. Simplifying Hypothesis

In order to study the weak nonlinearities of a MOSFET in saturation, let us consider the equivalent circuit of the intrinsic transistor depicted in Figure 3. The extrinsic part as well as the non-quasi static effects will be introduced in Section 4.1. The simplicity of this circuit allows us to establish compact analytical expression for the distortion figures of merit. The drain current I_{ds} is represented by a third order polynomial:

$$I_{ds} = g_{m1}V_{gs} + g_{m2}V_{gs}^2 + g_{m3}V_{gs}^3 + g_{d1}V_{ds} + g_{d2}V_{ds}^2 + g_{d3}V_{ds}^3 \qquad (1)$$

where the g_{mi} and g_{di} coefficients ($i = 1, \ldots, 3$) are respectively given by $1/i! \cdot \partial^i I_{ds}/\partial V_{gs}^i$ and $1/i! \cdot \partial^i I_{ds}/\partial V_{ds}^i$. Note that the cross-derivatives are not taken

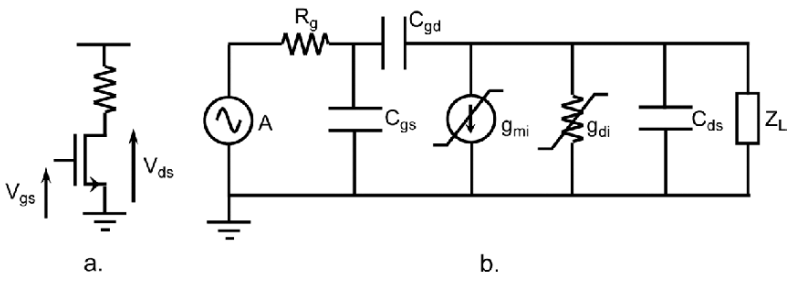


Figure 3. *a.* MOSFET in common-source configuration. *b.* Equivalent intrinsic quasi-static circuit of the transistor in *a*.

into account. Numerical simulations showed that the presence of the cross-terms improve the accuracy of the model, especially in the triode regime where they play a crucial role. We restrict us here to saturation.

The main capacitances are included in the model. As capacitive elements do not generate any significant harmonics in saturation, only the linearized capacitances are introduced. Only a few percents error is introduced by this simplification in the frequency band below 25 GHz [25]. Nevertheless, it is important to introduce these capacitances since C_{gd} influences the harmonics by the feedback, C_{ds} reduces the output impedance at high frequency, and C_{gs} filters the inputs. Moreover, the relative importance of C_{gd} increases from a technological generation to the next one. Compared to BJTs, MOSFETs feature more linear capacitances, which is an advantage of this technology [26].

3.3. Harmonic Distortion

The long-channel devices distortion is dominated by the transconductance distortion, since the output conductance g_d is almost bias independent in saturation. In that case, at low frequencies when the capacitances are neglected, the Taylor approach is used, and the harmonic distortion (HD) of order two and three (HD_2 and HD_3, respectively) are given by

$$HD_2 = \frac{A}{2}\left|\frac{g_{m2}}{g_{m1}}\right|, \quad HD_3 = \frac{A^2}{4}\left|\frac{g_{m3}}{g_{m1}}\right| \tag{2}$$

where A is the excitation amplitude. From these very simple expressions, it can already be concluded that a minimum of HD_3 exists around the threshold voltage, independently of short-channel effects. Indeed, this corresponds to the inflection point of the $I_{ds}(V_{gs})$ curve and it can be physically interpreted as follows. In weak inversion, the drain current has an exponential type behavior with respect to the gate voltage, as diffusion mechanism dominates. The g_{m3} coefficient is positive in this region and as a consequence, the device experiences gain expansion. In the strong inversion region, however, the mechanism for drain current is mainly due to drift and the drain current characteristics follows ideally a square-law (in reality, short-channel effects and mobility reduction affect the current shape). As a result, g_{m3} is nonzero and negative in strong inversion. The device experiences then gain compression. An important consequence of this feature is that g_{m3} passes through zero in the moderate inversion region. At this point, the device acts as an ideal square-law device and does not experience any third-order distortion. Higher order distortion components nevertheless also influence the third-order distortion, and even if HD_3 differs from zero at that bias, it experiences a minimum.

In submicron MOSFETs, the output conductance cannot be simply modeled by a constant Early voltage. This affects the HD of the circuit in Figure 3 as follows:

$$HD_2 \approx \frac{A}{2}\left|\frac{g_{m2}}{g_{m1}} - A_{vDC}\frac{g_{d2}}{Y_L + g_{d1}}\right|,$$
$$HD_3 \approx \frac{A^2}{4}\left|\frac{g_{m3}}{g_{m1}} - A_{vDC}^2\frac{g_{d3}}{Y_L + g_{d1}}\right| \tag{3}$$

where $Y_L = 1/Z_L$ is the load admittance and $A_{vDC} = -g_{m1}/(Y_L + g_{d1})$ is the linear DC voltage gain. The relations (3) confirm the intuition that the lower the load impedance is, the lower the effect of the g_d on nonlinearity is.

Even if the Taylor approach is commonly used, it suffers from two main drawbacks. First, it requires either the evaluation of the *I-V* derivatives that may become tricky when measurements data are used, or the addition of specific measurements. Second, it is limited to weak nonlinear systems. The integral function method was recently introduced to avoid these drawbacks [27, 28].

A comparison of the results obtained by relations (2) and (3) with measurements performed with LSNA at 900 MHz in a 50 Ω system showed that the distortion is dominated by the current-voltage characteristics (Figure 4). The static *I-V* characteristics are then sufficient to evaluate the distortion of a device up to a certain frequency. An analytical Volterra model confirms this interesting property, as it will be explained in the following.

Using the method of nonlinear currents applied to the circuit in Figure 3, the Volterra kernels are calculated. In this method, each nonlinear element of

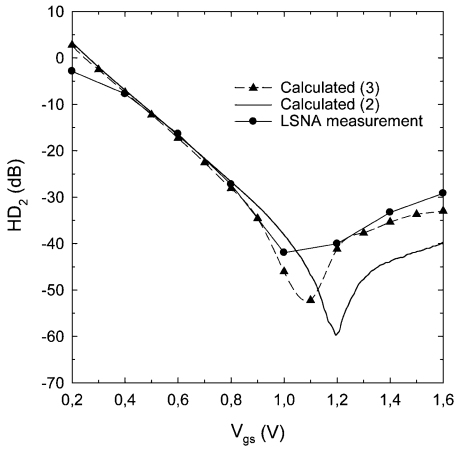


Figure 4. Harmonic distortion of order 2 evaluated by relations (2) and (3), compared to LSNA measurements at 900 MHz. PD FB MOSFET; 12 × 6.6 µm/0.25 µm; Threshold voltage V_{th} is 0.54 V; $V_{ds} = 1.2$ V; $A = 0.2$ V.

the circuit is converted into a linear element in parallel with current sources that represent the nonlinearity of this element. The current at each order of nonlinearity depends on the element voltages at all lower orders, in such a way that the currents may be calculated recursively.

From the evaluation of the Volterra kernels in the low GHz frequency range, the frequency variation of the HD figure-of-merit is accurately represented by a pole-zero [25] (see inset of Figure 5):

$$HD_2 \approx HD_{2DC} \left| \frac{1 + jf/f_{zHD_2}}{1 + jf/f_{pHD_2}} \right| \tag{4}$$

with

$$HD_{2DC} = \frac{A}{2} \left| \frac{g_{m2}}{g_{m1}} - A_{vDC} \frac{g_{d2}}{Y'_L} \right|, \tag{5}$$

$$f_{zHD_2} = \frac{1}{4\pi} \left[R_g (C_{gd} + C_{gs}) + \frac{g_{m2} Y'_L (C_{ds} + C_{gd}) - g_{d2} g_{m1} C_{gd}}{g_{m2} Y'^2_L + g_{d2} g^2_{m1}} \right]^{-1}, \tag{6}$$

$$f_{pHD_2} = \frac{1}{2\pi} \left[3R_g \left(C_{gd} \left(\frac{g_{m1}}{Y'_L} + 1 \right) + C_{gs} \right) + 3\frac{C_{ds} + C_{gd}}{Y'_L} - \frac{C_{gd}}{g_{m1}} \right]^{-1}, \tag{7}$$

$$Y'_L = g_{d1} + \frac{1}{Z_L} \tag{8}$$

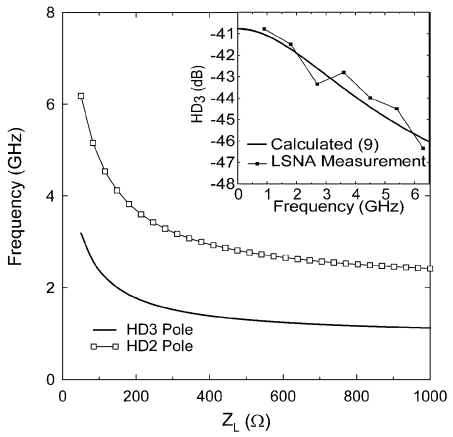


Figure 5. Dominant poles of HD. PD transistor, $12 \times 6.6\,\mu\text{m}/0.25\,\mu\text{m}$; $V_{ds} = V_{gs} = 1.2\,\text{V}$. Inset: frequency evolution of H_{D3} for the same device: Eq. (9) *versus* LSNA measurements; $A = 0.3\,\text{V}$.

and

$$HD_3 \approx HD_{3DC} \left| \frac{(1 + jf/f_{z1HD_3})(1 + jf/f_{z2HD_3})}{(1 + jf/f_{p1HD_3})(1 + jf/f_{p2HD_3})} \right| \tag{9}$$

with

$$HD_{3DC} = \frac{A^2}{4} \left| \frac{g_{m3}}{g_{m1}} + \frac{2g_{d2}g_{m2}}{Y_L'^2} + \frac{g_{m1}^2}{Y_L'^2} \left(\frac{2g_{d2}^2}{Y_L'^2} - \frac{g_{d3}}{Y_L'} \right) \right|, \tag{10}$$

$$f_{z1HD_3} = \frac{2g_{d2}^2 - Y_L' g_{d3}}{4\pi \left[\left(a g_{d3} - 2R_g g_{d2}^2 \right) C_{gd} + g_{d3} C_{ds} + \left(g_{d3} Y_L' R_g - 2R_g g_{d2}^2 \right) C_{gs} \right]}, \tag{11}$$

$$f_{z2HD_3} = \frac{1}{6\pi R_g C_{gd}}, \tag{12}$$

$$f_{p1HD_3} = \frac{-Y_L'/2\pi}{7C_{gd}\left(a - \left(Y_L'/7g_{m1}\right)\right) + 4(C_{ds} + Y_L' R_g C_{gs})}, \tag{13}$$

$$f_{p2HD_3} = \frac{-g_{m1}}{2\pi C_{gd}}, \tag{14}$$

$$a = 1 + R_g(g_{m1} + Y_L'). \tag{15}$$

From the Eq. (10), it is evident that the third-order distortion is not only generated by the third-order nonlinear coefficient g_{m3}, but also by the combination of lower-order terms.

The poles and zeros values depend on the bias point of the MOSFET and on the load impedance. It was found [25] that for a 0.25 µm SOI technology and for the bias of interest, the zeros in relations (4) and (9) lie at higher frequency than the poles. Furthermore, in Eq. (9), the dominant pole and zero of HD_3 are respectively f_{p1HD_3} and f_{z1HD_3}. In other words, both HD_2 and HD_3 are almost constant until the frequency f_{pHD_2} and f_{p1HD_3} is respectively reached. The dominant poles of a 0.25 µm PD MOSFETs, calculated by Eqs. (7) and (13), are plotted in Figure 5. If the load of the 0.25 µm PD transistor is for instance 200 Ω, the dominant pole of HD_3 is around 4.5 GHz.

It is interesting to note that the ratio between the pole of the voltage gain $A_v(f)$ and the dominant pole of $HD_2(HD_3)$ is almost constant at relatively high value of Z_L. Furthermore, it can be concluded from relations (7) and (13) that

$$\left| f_{pHD_2} \right| \geq \frac{\left| f_{pAv} \right|}{3}, \quad \left| f_{p1HD_3} \right| \geq \frac{\left| f_{pAv} \right|}{7} \tag{16}$$

where f_{pAv} is the dominant pole frequency of the voltage gain. Relations (16) give us a rule of thumb to find the frequency at which the low-frequency analysis (e.g., Taylor approach) is not valid anymore.

3.4. Intermodulation Distortion

A two-tone analysis is performed using the Volterra nonlinear current method. In order to simplify the expressions, C_{gd} was neglected i.e., a perfect feedback isolation is supposed. It was further assumed that the tones are close together with regard to the operation angular frequency ω, and that the global output admittance $G_0(\omega)$ verifies $G_0^*(\omega) = G_0(-\omega)$. This hypothesis is verified in our case (Figure 3) since $G_0(\omega) = Y_L + g_{d1}(\omega) + j\omega C_{ds}$ is linearly related to the frequency. Then, the intermodulation distortion of order 3 is given by:

$$IMD_3 = \frac{3}{4}A^2 \frac{1}{1+\omega^2 R_g^2 C_{gs}^2} |IM_3| \tag{17}$$

where

$$\begin{aligned} IM_3 = {} & \frac{g_{m3}}{g_{m1}} - \left(\frac{g_{m1}}{G_0(\omega)}\right)^2 \frac{g_{d3}}{G_0^*(\omega)} \\ & + \frac{2}{3} g_{d2} g_{m2} \left(\frac{1}{G_0(\omega) G_0^*(2\omega)} + \frac{1}{G_0(\omega) G_0^*(\Delta\omega)}\right) \\ & + \frac{2}{3} \left(\frac{g_{m1}}{G_0(\omega)}\right)^2 \frac{g_{d2}^2}{G_0^*(\omega)} \left(\frac{1}{G_0(2\omega)} + \frac{1}{G_0^*(\Delta\omega)}\right). \end{aligned} \tag{18}$$

It follows from these equations that the IMD_3 not only depends on the third-order nonlinearity of the transconductance g_{m3}, but also on the output conductance at low-frequency $G_0(\Delta\omega)$ and at the second harmonic $G_0(2\omega)$. Also, the only action of C_{gs} is a lowering of the signal magnitude at the input of the device. As the output admittance of a PD SOI MOSFET is frequency dependent, we expect from those last relations to measure a variation of IMD_3 as a function of the tones separation for that device [29]. This will be experimentally discussed in Section 5.

4. SOI MOSFET CAD Modeling

As discussed before, the analytical modeling exhibits some limitations with respect to the amplitude of the input signal and the degree of linearity of the

device. It is therefore not suited for circuit design through CAD tools. A nonlinear empirical modeling of SOI MOSFETs adapted to CAD will be presented in this section. The model is valid in both small and large signal regimes and from DC to RF. The aim of the approach is to provide a model very helpful for circuit design, whenever fast modeling is needed. The empirical approach is adopted for nonlinear current and capacitance modeling in order to reduce the extraction procedure. Furthermore, the determination of empirical model parameters does not need any knowledge about technological process as it is based on observation.

The dedicated electrical circuit model will be presented in Section 4.1. All the elements of the model are described in details in Section 4.2. Finally, the validity of the model is investigated through measurements, compared to simulation results. The extraction procedure of the model parameters is omitted and the reader must refer to reference [32]. Note that the model is implemented into the Advanced Design System (ADS, Agilent Technologies) environment, in which all simulations are performed.

4.1. Electrical Equivalent Circuit

In order to obtain an accurate model valid in a large frequency range, it is necessary to complete the simple model presented in Figure 3. On the other hand, the model has to be efficient whenever Harmonic Balance (HB) algorithm is used. For that purpose, two conditions have to be satisfied: first, the charge conservation principle and second, the continuity and derivability of the nonlinear equations. The first point is easily achieved if the equations describing the nonlinear capacitances derive from charge equations. The second point is achieved if single and continuous nonlinear equations are used for the current and charge in the whole regime operation. In our approach, the drain and source charges are considered as linear and only the charge of the gate is taken as nonlinear.

Considering the above remarks on the charge and the current, the equivalent circuit shown in Figure 6 represents the SOI MOSFET electrically. The grey area denotes the intrinsic part of the device and the arrows show the nonlinear elements. The resistive elements R_i and R_{gd} account for the non-quasi static effects. The I_{kink} nonlinear current is specific to the floating body devices. In addition to the active part of the device, the model includes the extrinsic elements: R_g, R_d and R_s are respectively the gate, the drain and the source resistances, while L_g, L_d and L_s are the access inductances and finally, C_{pd} and C_{pg} are the extrinsic (and access) parasitic capacitances. In order to maintain the extraction procedure as simple and as quick to implement as possible, this model is very compact. As a matter of fact, the overlap and fringing capacitances

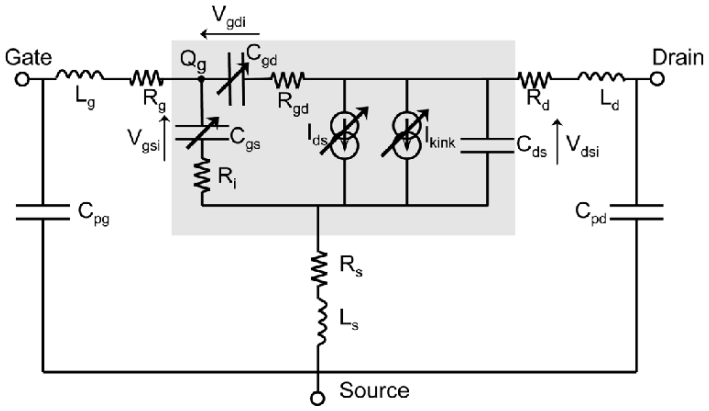


Figure 6. Electrical equivalent circuit used for modeling the SOI MOSFET. The grey area denotes the intrinsic part of the transistor. The arrows show the non linear elements.

between gate and source and drain and source are implicitly included in the intrinsic C_{gs} and C_{gd}.

4.2. Nonlinear Drain Current Equation

The drain to source current is modeled by the empirical equation (19), which was based on the Angelov's model [30] (see Chapter 5). Some modifications [31, 32] were undertaken in order to better describe the current of SOI MOSFETs.

$$\begin{aligned} I_{ds} &= I_{pk}\{1 + P_{ol1}(\tanh(\Psi))\}\, P_{ol2} \tanh\left\{\left(\alpha_1 + \alpha_2 V_{gsi}\right) V_{dsi}\right\} \\ \Psi &= P_1\left(V_{gsi} - V_{pk}\right) + P_2\left(V_{gsi} - V_{pk}\right)^2 + P_3\left(V_{gsi} - V_{pk}\right)^3 \\ P_{ol1} &= K_0 + K_1 V_{gsi} + K_2 V_{gsi}^2 + K_3 V_{gsi}^3 \\ P_{ol2} &= 1 + \lambda_1 V_{gdi} + \lambda_2 V_{gdi}^2 + \lambda_3 V_{gdi}^3 \end{aligned} \tag{19}$$

where V_{gsi} and V_{dsi} are respectively the intrinsic gate to source and drain-to-source potentials, and V_{gdi} is the intrinsic gate-to-drain potential; $I_{pk}, V_{pk}, \alpha_i, \lambda_i, P_i$ $(i = 1, \ldots, 3)$ and K_i $(i = 0, \ldots, 3)$ are the fitting parameters of the model.

The $(1 + P_{ol1}\tanh(\Psi))$ term describes the current control by the gate voltage. The polynomial P_{ol2} was developed to model the saturation regime of deep sub micron MOSFETs. The term $\tanh((\alpha_1 + \alpha_2 V_{gs})V_{ds})$ describes the linear zone and the transition between the linear regime and saturation. Note that the Eq. (19) describes the DC drain current of FD or PD devices with body ties. Section 4.3 is dedicated to the kink effect modeling in FB devices.

4.3. Kink Effect Modeling

In the case of a FB device, the kink effect is modeled by a large signal current that is frequency dependent. We consider that two currents constitute the total drain-to-source current: the kink-free DC channel current I_{ds} (Eq. (19)) and a DC current I_{kink} specific to the kink. Thus, the total DC drain to source current I_{dsT} is given by:

$$I_{dsT} = I_{ds} + I_{\text{kink}} \tag{20}$$

The current I_{kink} is given by the empirical Eq. (21) [32, 33], and was obtained from experimental data on FB and BT FETs and Eq. (20).

$$I_{\text{kink}} = I_{ks} V_{gsi} V_{dsi} (1 + cV_{dsi}) \times \left\{ 1 + \tanh\left(a \left(V_{dsi} - \frac{b}{\sqrt{V_{gsi} + V_{th}}} \right) \right) \right\} \tag{21}$$

where V_{th} is the measured threshold voltage at low drain current and the other symbols (I_{ks}, a, b, c) are the model fitting parameters. Figure 7a shows the measured current, the simulated kink-free current I_{ds} and the simulated total current I_{dsT}, while Figure 7b shows the DC I_{kink} simulated by Eq. (21) after extraction. The I_{ds} parameters are first extracted in the pre-kink region (i.e., $V_{ds} < 0.6\,\text{V}$ in the example). The I_{kink} parameters are next extracted in all regime operation.

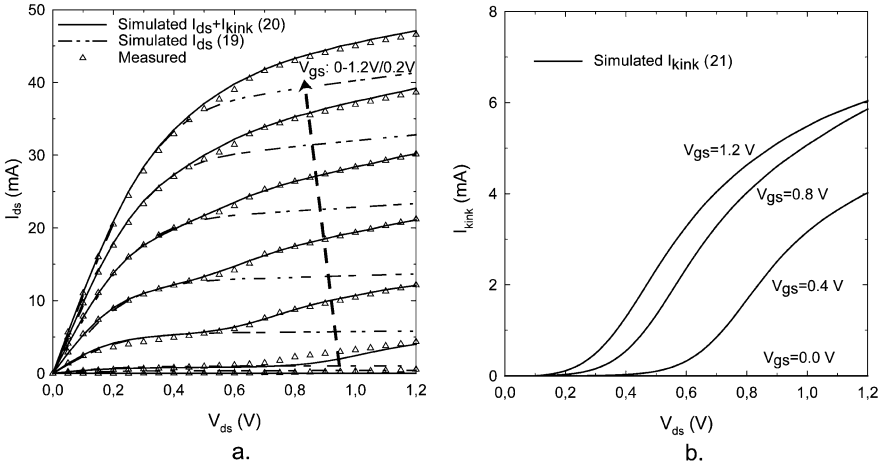


Figure 7. *a.* Measured drain current (symbols), simulated kink-free (doted line) and simulated total drain current (solid line) of a PD FB device *versus* the drain bias ($60 \times 1\mu\text{m}/0.12\mu\text{m}$). *b.* Simulated kink current.

Note that Eq. (21) does not give any frequency dependence. In order to include a frequency dispersion model, Eq. (21) is expressed in the frequency domain and it is balanced by a frequency dispersion function $f_{dm}(\omega) = e^{-|\omega|\tau_k}$:

$$I'_{\text{kink}}(\omega) = e^{-|\omega|\tau_k} \int_{-\infty}^{\infty} I_{\text{kink}}(t)\, e^{-j2\pi ft} dt \tag{22}$$

where τ_k is a time constant defining the cut-off frequency of the kink.

The empirical equation $f_{dm}(\omega)$ was obtained by low frequency measurements of the drain conductance. The frequency dependent large signal current $I'_{\text{kink}}(\omega)$ given by (22) can be expressed in the time domain by applying the inverse Fourier transform. Hence, Eq. (22) becomes:

$$I'_{\text{kink}}(t) = f_{dm}(t) \otimes I_{\text{kink}}(t) \tag{23}$$

where $f_{dm}(t)$ is the frequency dispersion model calculated in the time domain. Its analytical expression is given by:

$$f_{dm}(t) = F^{-1}(f_{dm}(\omega)) = \frac{\tau_k}{\pi\left(\tau_k^2 + t^2\right)} \tag{24}$$

To understand how this model acts, let us consider the time domain periodic kink current $I_{\text{kink}}(t)$ in the case of a single tone excitation at angular frequency ω_0. From Eq. (23), it can be easily shown that

$$I'_{\text{kink}}(t) = \sum_{n=-\infty}^{\infty} e^{-|n|\omega_0\tau_k} I_{\text{kink n}} e^{j\omega_0 t} \tag{25}$$

with

$$I_{\text{kink n}} = \frac{1}{T_0} \int_0^{T_0} I_{\text{kink}}(t)\, e^{-j2\pi ft} dt, \quad \text{and} \quad T_0 = \frac{2\pi}{\omega_0} \tag{26}$$

These equations show that $f_{dm}(\omega)$ balances the magnitude of each Fourier current coefficient $I_{\text{kink n}}$ at each frequency component $n\omega_0$. Thus, the large signal kink current is reduced as the frequency increases.

The kink effect in the time domain is depicted in Figure 8a, where the simulated instantaneous large signal current $I_{dsT}(t)$ is plotted *versus* the instantaneous drain voltage $V_{ds}(t)$. For that simulation, we applied a large signal on the drain of the device (V_{ds} swing is 0.5 V) and the gate to source voltage variations were suppressed using a DC-block capacitance. We carried out the simulation for various frequencies between 10 Hz and 100 MHz. At low frequencies the large signal current includes the kink effect, while at high frequency it disappears.

In Figure 8b, the small signal drain conductance g_{d1} is simulated as a function of the drain bias in the inversion regime, for frequencies varying from

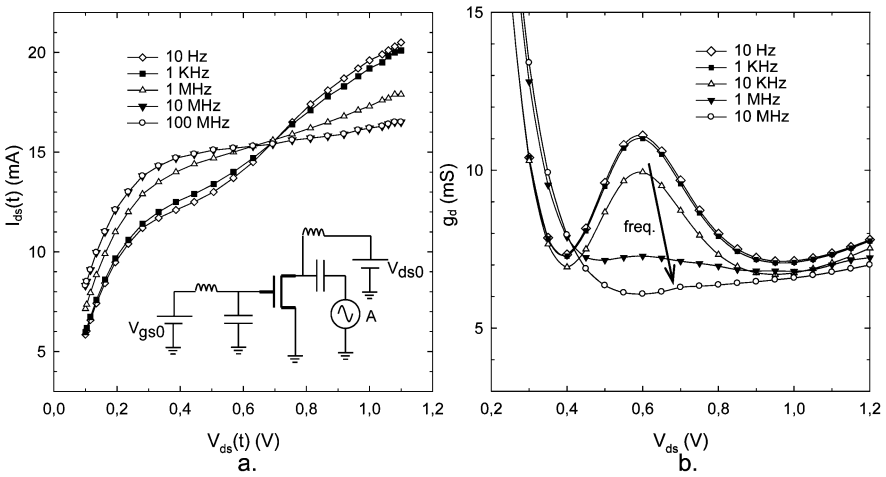


Figure 8. *a.* Simulated instantaneous large signal drain current versus the instantaneous drain voltage for various applied frequencies. $V_{ds0} = 0.6\,\text{V}$, $V_{gs0} = 0.6\,\text{V}$, $A = 0.5\,\text{V}$. Inset: schematic test bench used for simulation. *b.* Simulated small signal drain conductance versus the drain bias for various frequencies.

10 Hz to 10 MHz. We observe discrepancies with measured results shown in Figure 2b. Nevertheless, the kink phenomenon is well reproduced, as well as its vanishing above 1 MHz.

This technique (Eq. (22) or (23)) allows modeling easily a nonlinear element with frequency dispersion due to charge accumulation, without introducing additional electrical elements (capacitances) in the equivalent circuit. This implies that additional test elements and time consuming measurements for extraction can be avoided.

4.4. Charge Modeling

Following the assumptions in the introduction of Section 4, a single gate charge model Q_g is used for calculating the nonlinear intrinsic capacitances C_{gs} and C_{gd}. Because the charge is not directly measurable, it is necessary to carry out physics based simulation in order to obtain charge data. Accurate results are easily obtained by drift-diffusion simulation, using for example the ATLAS-SILVACO simulator. From that data, the following charge equation was elaborated [32]:

$$Q_g\left(V_{gs}, V_{ds}\right) = C_0\left(L_{\text{gate}} - 2L_m\right) W_f n_f \left\{C_{gsf}\left(V_{gs}, V_{ds}\right) + C_{gdf}\left(V_{gs}, V_{ds}\right)\right\} + Q_0,$$

$$C_{gsf}\left(V_{gs}, V_{ds}\right) = \left(C_{gg0} + \tanh\left(\frac{V_{ds}^2}{\gamma V_{gs}^2}\right)\right) \times \left(C_{gg1}V_{gs} + C_{gg2}V_{gs}^2 + C_{gg3}V_{gs}^3\right),$$

$$C_{gdf}\left(V_{gs}, V_{ds}\right) = \left(C_{gd0} + \tanh\left(-\frac{V_{gs}^2}{V_\alpha}\right)\right)\left(C_{gd1}V_{gd} + C_{gd2}V_{gd}^2\right). \quad (27)$$

The capacitances C_{gs} and C_{gd} are calculated by combining (27) to (28) and (29).

$$C_{gs}\left(V_{gs}, V_{ds}\right) = \left.\frac{\partial Q_g}{\partial v_{gs}}\right|_{v_{ds}=cte} + \left.\frac{\partial Q_g}{\partial v_{ds}}\right|_{v_{gs}=cte}, \quad (28)$$

$$C_{gd}\left(V_{gs}, V_{ds}\right) = -\left.\frac{\partial Q_g}{\partial v_{ds}}\right|_{v_{gs}=cte} \quad (29)$$

where L_{gate}, W_f and n_f are respectively the gate length, the gate width per finger and the number of fingers; $C_0, L_m, \gamma, V_\alpha$ and C_{ggi}, C_{gdi} $(i = 1, \ldots, 3)$ are the charge model fitting parameters; and Q_0 is the depletion charge in the gate terminal at $V_{gs} = V_{ds} = 0\,\text{V}$.

4.5. Scaling Rules

Empirical modeling is not appropriate for including scaling rules as a function of the gate length. However, most of the model parameters follow simple scaling rules as a function of the gate width and the number of fingers. The scaling rules of the different model parameters are summarized in Table 1.

4.6. Model Validity

The validity of a model needs to be demonstrated through measurements in large operation conditions. However, the model verification should be limited

Table 1. Scaling rules as a function of the total gate width W_t ($W_t = W_f n_f$) and the number of fingers n_f.

Parameter	Scaling Rule
$C_{pd}, C_{pg}, C_{gs}, C_{gd}, C_{ds}, I_{ds}, I_{\text{kink}}$	W_t
R_s, R_d	$1/W_t$
R_g	W_t/n_f^2

with regard to its application. For example, the aim of this chapter is to study the nonlinearities of SOI MOSFETs; for that reason we verify that distortion and intermodulation are well reproduced by the model. The validity of the model in small signal operation has been shown elsewere [32].

Figure 7 shows that the DC drain current of devices with and without FB is accurately described. But, the designer must be aware that the current model is not accurate in sub-threshold bias operation. This is not a problem, as in MMIC design the transistors are commonly biased in the strong inversion regime to reach a high operation frequency.

The nonlinearity of the device at high frequency is also correctly described. Indeed, a good agreement between measurements and simulations is observed in Figure 9a for the output power at the fundamental frequency, the second and the third harmonics for all DC gate biases when a single tone power is applied on the gate of the device. The result of a two-tones test is shown in Figure 9b. In that case again, we note an excellent prediction of the nonlinear behaviour of the device.

5. SOI MOSFET Linearity

In this section, the linearity properties of PD FB and BT devices, as well as FD MOSFETs are investigated through simulation and measurements. The simulations are performed in the ADS simulator with the model described in Section 4 using the Harmonic Balance algorithm, and the results are interpreted

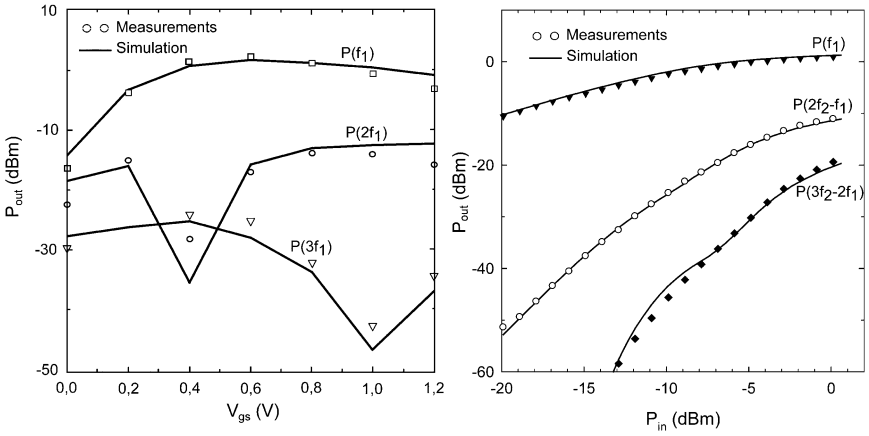


Figure 9. *a.* Output power of the fundamental, the second and the harmonic versus the gate bias. $V_{ds} = 0.6\,\mathrm{V}$, $f_1 = 2\,\mathrm{GHz}$, $P_{in} = -5\,\mathrm{dBm}$. *b.* Output power of the fundamental frequency, the 3rd order and the 5th order intermodulation products. $f_1 = 2\,\mathrm{GHz}$, $f_2 = 2.001\,\mathrm{GHz}$ $V_{gs} = 0.6\,\mathrm{V}$, $V_{ds} = 0.6\,\mathrm{V}$. $R_{load} = 50\,\Omega$ ($60 \times 1\,\mu\mathrm{m}/0.12\,\mu\mathrm{m}$).

by the analysis presented in Section 2.2. All measurements were performed on-wafer with the MOSFET source and back-gate grounded.

5.1. Harmonic Distortion

Even if the FD transistor has a higher current drive and transconductance than a PD transistor with the same technological parameters (cfr. Section 2), the harmonic distortion factors of merit are not affected since they are defined from the *ratios* of the g_{mi} and g_{di}, as explained before (Eq. (3)). This is depicted in Figure 10a. Care must be undertaken while the evaluation of the g_{di} coefficients in the case of FB devices. Indeed, when these coefficients are extracted at DC or at very low frequency, and because of the kink effect, these coefficients nullify at some bias inducing sweet spots. Nevertheless, the kink effect disappearing at high frequency, these biases are not interesting for improving the linearity of MMICs. In order to use the analytical formulation presented in Section 2.2, the coefficients g_{di} of any FB device should thus be evaluated at the working frequency.

The comparison between PD FB and BT transistors in 0.12 µm technology is performed in Figure 10b. The total harmonic distortion (THD) of the BT device is slightly higher (about 3 dB), over the entire frequency range. This is

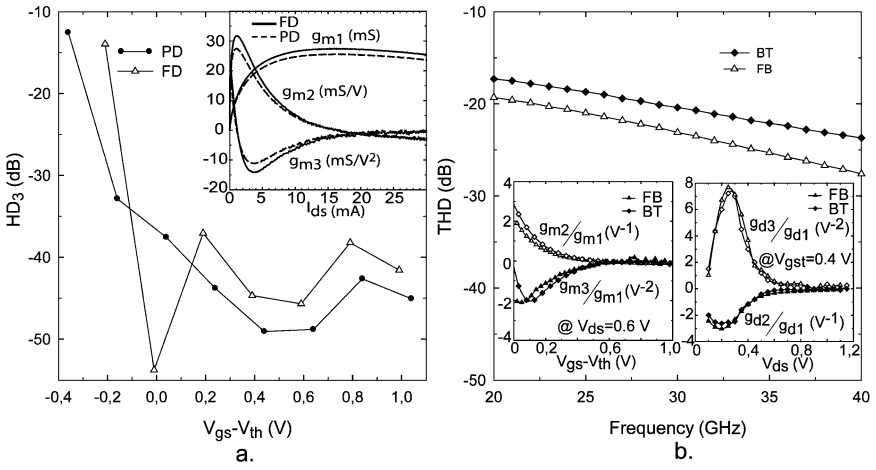


Figure 10. *a.* Measured HD_3 *versus* the gate voltage overdrive V_{gst} of PD and FD devices ($12 \times 6.6\,\mu\text{m}/0.25\,\mu\text{m}$). $f_0 = 900\,\text{MHz}$, $V_{ds} = 1.2\,\text{V}$; $A = 0.2\,\text{V}$; Inset: g_{mi} coefficient of these two transistors; $V_{ds} = 1.2\,\text{V}$. *b.* Simulated Total Harmonic Distortion of PD FB and BT ($60 \times 1\,\mu\text{m}/0.12\,\mu\text{m}$) transistors *versus* applied frequency. $P_{in} = 0\,\text{dBm}$, $V_{gs} = 0.6\,\text{V}$, $V_{ds} = 0.6\,\text{V}$. First inset: $g_{mi}/g_{m1} (i = 2, 3)$ ratios of both devices *versus* the gate voltage overdrive V_{gst}. $f = 20\,\text{GHz}$. Second inset: g_{di}/g_{d1} $(i = 2, 3)$ ratios of both devices *versus* V_{ds}. $f = 20\,\text{GHz}$.

explained by somewhat lower ratios of g_{mi}/g_{m1} and g_{di}/g_{d1} $(i = 1, 2)$ for the FB device (insets of Figure 10b). Moreover, the kink effect in the FB device is totally suppressed in the g_{d2}/g_{d1} ratio at high frequencies. From these results, it can be concluded that HD is not much affected by the type of device considered. This is due to the similar shape of drain current, which is the main source of nonlinearity, at high frequency where the kink vanishes.

5.2. Intermodulation Distortion

Even though the FB effect does not impact highly the harmonic distortion, it is expected from Section 3.4 to affect the intermodulation ditortion, even at RF. Indeed, Eqs. (17) and (18) show that IMD_3 is not only dependent on $G_0(\omega)$ but also on $G_0(\Delta\omega)$. The G_0 dependence on frequency was depicted in Figure 2. This is confirmed by two-tones tests performed on FD and PD devices. Measurements indicate that IMD_3 of the PD FB device is dependent on the tones separation Δf for bias points where the *I–V* exhibits a kink (Figure 11a).

The model presented in Section 4.4 was used to simulate the third order output intercept point (OIP_3) as a function of the two tones separation and the load resistance. The simulation results show that the intermodulation properties of the FB device are modified when the frequency spacing value is inferior to the kink cut-off frequency, which is around 1 MHz in the example (Figure 11b). This transition is obvious for of a high load resistance value, while in the case of a low load resistance, the Δf influence is quasi negligible. For low load values,

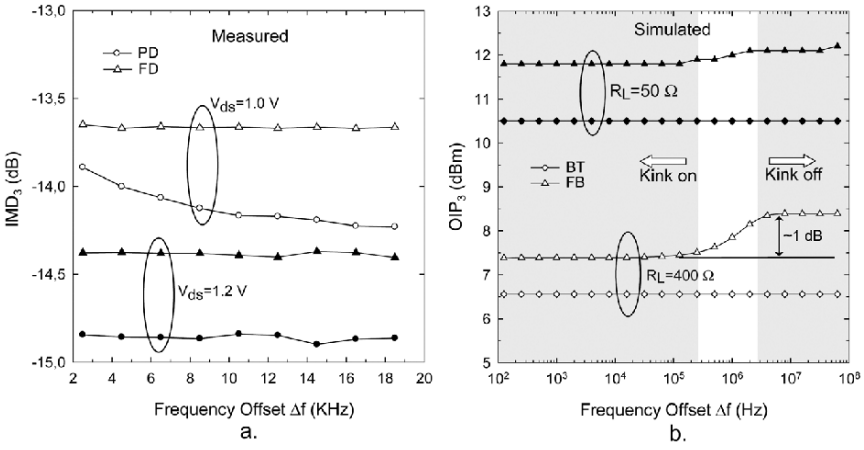


Figure 11. *a.* IMD_3 measured by LSNA at 900 MHz; $V_{gs} = 1.0$ V; $A = 0.6$ V. *b.* Simulated OIP_3 as a function of the frequency offset Δf and two different load values. $V_{ds} = 0.6$ V, $V_{gs} = 0.6$ V.

the linearity of the device is mainly influenced by the transconductance g_m, while as the load increases the drain conductance becomes more significant, as expected from relation (18). This remark also holds for HD, as seen in Eq. (3). The OIP_3 of the BT kink free device is totally independent on the frequency spacing. The kink effect acts thus as a low frequency memory effect on the intermodulation properties of a floating body device.

6. Summary

An empirical nonlinear model for SOI devices was presented. The empirical approach is useful when a rapid extraction is required; however, it is not physically scalable along the gate length dimension. This model is well suited to describe the behavior of SOI transistors along the frequencies under small and large signal operation. The floating body effects were modeled by an independent current source that takes the low frequency dispersion into account. The model was used to calculate different nonlinear figures of merit. The simulations are confronted to measurements and explained by a simple analytical model, based on the Volterra series approach. The analytical formulation permits to determine easily the frequency behavior of these nonlinear figures of merit. It was also explained that at high frequency, the FB effects do affect the IMD, but not the HD.

Acknowledgments

This work would not be realized without Professors. J.-P. Raskin (UCL), F. Danneville (IEMN) and G. Dambrinne (IEMN) who advised it; the authors kindly thank them. The authors are also grateful to Prof. Dominique Schreurs (Katholike Universiteit Leuven) for her help in the measurements as well as for the outstanding discussions. A special thanks goes to Pascal Simon and Sylvie Lepilliet for the experiments. The authors would like to thank Dr. Franz Sischka (Agilent Technologies) for his continual help on ICCAP.

References

[1] Annema, A.-J.; Nauta, B.; van Langevelde, R.; Tuinhout, H. “Analog circuits in ultra-deep-submicron CMOS”, *IEEE J. Solid-State Circuits*, **2005**, *40(1)*, 132–143.

[2] Colinge, J.-P. “Fully-depleted SOI CMOS for analog applications”, *IEEE Trans. Electron Dev.*, **1998**, *45(5)*, 1010–1016.

[3] Flandre, D. *et al.* “Fully depleted SOI CMOS technology for heterogeneous micropower, high-temperature or RF microsystems”, *Solid-State Electron.*, **2001**, *45(4)*, 541–549.

[4] Subramanian, V. *et al.* "Device and circuit-level analog performance trade-offs: A comparative study of bulk- versus Fin-FETs", *IEEE Int. Electr. Dev. Meeting*, Washington, USA, **December 2005**.

[5] Raskin, J.-P.; Viviani, A.; Flandre, D.; Colinge, J.-P. "Substrate crosstalk reduction using SOI technology", *IEEE Trans. on Electr. Dev.*, **1997**, *44(12)*, 2252–2261.

[6] Raynaud, C. *et al*. "Is SOI CMOS a promising technology for SOCs in high frequency range?" *207th Electrochemical Society Meeting, Proc. Silicon Insulator Technol. Dev.*, Quebec City, Canada, **May 2005**, 331–344.

[7] Lee, M.S.L.; Tenbroek, B.M.; Redman-White, W.; Benson, J.; Uren, M.J. "A physically based compact model of partially depleted SOI MOSFETs for analog circuit simulation", *IEEE J. Solid-State Circuits*, **2001**, *36(1)*, 110–121.

[8] Iniguez, B.; Ferreira, L.; Gentinne, B.; Flandre, D. "A physically-based C_∞-continuous fully-depleted SOI MOSFET model for analog applications", *IEEE Trans. Electron Dev.*, **1996**, *43(4)*, 568–575.

[9] Veeraraghavan, S.; Fossum, J.G. "A physical short-channel model for the thin-film SOI MOSFET applicable to device and circuit CAD", *IEEE Trans. Electron Dev.*, **1988**, *35(11)*, 1866–1875.

[10] Colinge, J.-P. *Silicon-on-Insulator Technology: Materials to VLSI*, Kluwer Academic Publisher, **1991**.

[11] Howes, R.; Redman-White, W. "A small-signal model for the frequency-dependent drain admittance in floating-substrate MOSFET's", *IEEE J. Solid-State Circuits*, **1992**, *27(8)*, 1186–1193.

[12] Chen, J.; Fang, P.; Ko, P.K.; Hu, C.; Solomon, R.; Chan, T.-Y.; Sodini, C.G. "Noise over-shoot at drain current kink in SOI MOSFET", *Proc. IEEE Int. SOI Conf.*, **1990**, 40–41.

[13] Kilchytska, V.; Levacq, D.; Vancaillie, L.; Flandre, D. "On the great potential of non-doped MOSFETs for analog applications in partially-depleted SOI CMOS process", *Solid-State Electron.*, **2005**, *49(5)*, 708–715.

[14] Tseng, Y.-C. *et al.* "AC floating body effects and the resultant analog circuit issues in submicron floating body and body-grounded SOI MOSFET's", *IEEE Trans. Electron Dev.*, **1999**, *46(8)*, 1685–1692.

[15] Joshi, R.V. *et al.* "Effects of gate-to-body tunneling current on PD/SOI CMOS SRAM", *Digest of Techn. Papers. 2001 Symposium on VLSI Technology*, **2001**, 75–76.

[16] Mercha, A.; Rafi, J.M.; Simoen, E.; Augendre, E.; Claeys, C. "Linear kink effect induced by electron valence band tunneling in ultrathin gate oxide bulk and SOI MOSFETS", *IEEE Trans. on Electr. Dev.*, **2003**, *50(7)*, 1675–1682.

[17] Lederer, D.; Flandre, D.; Raskin, J.-P. "AC behavior of gate-induced floating body effects in ultrathin oxide PD SOI MOSFETs", *IEEE Electr. Dev. Lett.*, **2004**, *25(2)*, 104–106.

[18] Vancaillie, L. *A Methodology for Characterizing and Introducing MOSFET Imperfections in Analog Top-Down Synthesis and Bottom-up Validation*, Ph.D. Thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium, **2005**, available from http://edoc.bib.ucl.ac.be/ucl/.

[19] Wambacq, P. *Distortion Analysis of Analog Integrated Circuits*. Dordrecht: Kluwer, **1998**.

[20] van Langevelde, R. *et al.* "RF-distortion in deep-submicron CMOS technologies", *Technical Digest of Int. Electron. Dev. Meeting*, **2000**, 807–810.

[21] Sansen, W. "Distortion in elementary transistor circuits", *IEEE Trans. Circ. & Syst. II: Analog and Digital Signal Process.*, **1999**, *46(3)*, 315–325.

[22] Bussgang, J.; Ehrman, L.; Graham, J. "Analysis of nonlinear systems with multiple inputs", *Proc. IEEE*, **1974**, *62(8)*, 1088–1119.

[23] Mass, S. "A general-purpose computer program for the volterra-series analysis of nonlinear microwave circuits", *Proc. IEEE Int. Sympos. Microwave Theory Techn.*, **1988**, 311–314.

[24] Verspecht, J.; Debie, P.; Barel, A.; Martens, J. "Accurate on wafer measurement of phase and amplitude of the spectral components of incident and scattered voltage waves at the signal ports of a nonlinear microwave device", *IEEE Int. Microwave Sympos. Digest*, **1995**, *3*, 1029–1032.

[25] Parvais, B.; Raskin, J.-P. "Analytical expressions for distortion of SOI MOSFETs using the volterra series", *Proc. 12th European GAAS*, Amsterdam, The Netherlands, **2004**, 223–226.

[26] Lee, T.-Y.; Cheng, Y. "High-frequency characterization and modeling of distortion behavior of MOSFETs for RF IC design", *IEEE J. Solid-State Circuits*, **2004**, *39(9)*, 1407–1414.

[27] Cerdeira, A.; Alemán, M.A.; Estrada, M.; Flandre, D. "Integral function method for determination of nonlinear harmonic distortion", *Solid-State Electron.*, **2004**, *48(12)*, 2225–2234.

[28] Parvais, B.; Raskin, J.-P.; Cerdeira, A.; Estrada, M. "Application of integral function method for distortion analysis of microwave transistors", *Asia Pacific Microwave Conf.*, New Delhi, India, **December 2004**.

[29] Adan, A.O.; Yoshimasu, T.; Shitara, S.; Tanba, N.; Fukurni, M. "Linearity and low-noise performance of SOI MOSFETs for RF applications", *IEEE Trans. Electron Dev.*, **2002**, *49(5)*, 881–888.

[30] Angelov, I.; Zirath, H.; Rorsman, N. "A new empirical nonlinear model for HEMT and MESFET devices", *IEEE Trans. Microwave Theory Techn.*, **1992**, *40(12)*, 2258–2266.

[31] Siligaris, A.; Vanmackelberg, M.; Dambrine, G.; Vellas, N.; Danneville, F. "A new empirical non-linear model for SOI MOSFET", *Proc. 10th European GAAS*, Milan, Italy, **2002**, 101–104.

[32] Siligaris, A. *Modélisation grand signal de MOSFET en hyperfrequences: application à l'étude des non linearités des filières* SOI, PhD Thesis, IEMN, Lille, France, **2004**; available from http://www.univ-lille1.fr/anodegroup/.

[33] Siligaris, A.; Dambrine, G.; Danneville, F. "Non-linear modeling of the kink effect in deep sub-micron SOI MOSFET", *Proc. 12th Eur. GAAS*, Amsterdam, The Netherlands, **2004**, 47–50,

Chapter 7

CIRCUIT LEVEL RF MODELING AND DESIGN

Nobuyuki Itoh
Semiconductor Company Toshiba Corporation, 2-5-1, Kasama, Sakae-ku, Yokohama, 247-8585, Japan
E-mail: nobuyuki3.ito@toshiba.co.jp

Abstract: The compact model has been improved due to device scaling and its accuracy has been going to be acceptable for analog circuit design. However, by viewpoint of RF circuit prediction, its accuracy is still poor even if using the recent MOSFET's compact model because it is necessary to implement all parasitic components effects to obtain good accuracy of RF circuit design. Moreover, it has still some insufficient phenomena in the recent small geometry MOSFET. One is the mobility degradation due to STI stress and another one is the channel noise enhancement due to hot carrier effects. This chapter focuses on and describes these uncovered or insufficient characterizations for MOSFET and their influence on RF design, especially voltage-controlled oscillator design.

Key words: RF Model; MOSFET; STI Stress; Scalable Parasitic Components Model; Channel Noise; Voltage-Controlled Oscillator Design

1. Introduction

The downscaling of the design rule of semiconductor technology has realized many features that were unavailable in previous generations of LSIs. Logic LSIs have come to employ higher-level integration and provide high-speed functions, leading to realization of high-performance CPUs such as Pentium-IV. RF-analog LSIs [1] also employ higher-frequency operation such as 5 GHz front-end and more than 10 GHz building block of radio communications. Since the cut-off frequency of MOSFET is approximately dependent on inverse gate length, cut-off frequency of over 200 GHz has already been achieved using sub 0.1 micron design rule. Although scaling-down is thought

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, 181–207.

not to pose problems, for either logic circuit or analog circuit design, there are drawbacks in the case that part of the circuit is analog. RF circuit designers are often faced with the inaccuracy and inconvenience of a compact model since the extraction of the parameter set is not perfect and sometimes the model does not express many physical phenomena even if it covers them.

The drawbacks are increased flicker noise due to introducing oxynitride in recent small geometry MOSFETs, current degradation due to shallow trench isolation (STI) stress, and increasing channel noise due to hot carrier effect in small geometry MOSFETs. Moreover, the normal MOSFET model is insufficient for RF circuit design since MOSFET's parasitic components are not introduced in the Spice model and the RF characteristics are significantly influenced by parasitic components [2].

The issue of STI stresses [3], basically, mechanical stress induced by STI, affects the carrier mobility and this influence depends on distance between edge of STI and channel. In the case of RF-analog circuit, multi-gate finger structure of MOSFET has been widely used, and MOSFET with such a structure has many channels. The mobility degradation due to STI stress differs for each channel. Therefore, drain current and transconductance of MOSFET are not precisely proportional to the number of gate fingers.

For RF design, substrate network model is one of the most important and it has often influenced circuit design accuracy, especially noise. There are some parasitic network models but there is no scalable parasitic network model. The scalable substrate model is strongly required by almost all designers.

Increases in channel noise due to scaling down of gate length is the most serious problem and has been the object of a greater deal of study [4–16] since this phenomenon had not been introduced in any Spice models although its existence has been known for over ten years.

This chapter, focuses on STI stress, parasitic component network and the channel noise enhancement of small geometry MOSFET and describe their influence on current mirror design and RF voltage-controlled oscillator design.

2. STI Stress

2.1. Origin of STI Stress

In the Si integrated circuit process, the isolation region was basically formed with thermal oxidation process. However, thermal oxidation process such as LOCOS isolation was limited so as to minimize isolation region due to bird's beak. Hence, shallow trench isolation (STI) has been utilized below $0.25\,\mu m$ process technologies. STI consists of shallow (approximately 0.3 to $0.5\,\mu m$) trench isolation etching and oxide is filled in it. These types of trench isolation had been used in BiCMOS process for over ten years. Of course, in the case

of BiCMOS process, approximately 5 μm depth deep trench isolation (DTI) or the combination of DTI and STI were used since the depth of collector buried layer is approximately 3 to 4 μm from surface. Also, mechanical stress of DTI or the combination of DTI and STI has been the object of much study [17, 18]. The determined stress in that work showed compressive stress, which was observed around trench isolations and it also depended on the distance from trench isolation. Figure 1 shows compression stress by DTI (left hand) and leakage current of pn-junction as a function of distance from trench isolation (right hand).

2.2. STI Stress on Small Geometry MOSFET

Similar to isolation in the case of BiCMOS process [19], the mechanical stress in the vicinity of active area (AA) is determined by the distance from STI edge. Thus, the mobility of electron of NMOSFET decreases as a function of inverse of the distance between them, resulting in -15% at the vicinity of STI edge. On the other hand, the mobility of hole of PMOSFET increases as a function of inverse of the distance between them, resulting in $+15\%$ at the vicinity of STI edge, vice versa.

These phenomena make it difficult to keep model scalability of MOSFET. In the conventional MOSFET's Spice model scalability of L_g and W_g is kept for both DC and CV parameters. Hence, DC characteristics depend on W_g/L_g but not on AA. However, it is necessary to add some parameter to correct this mobility dependence of distance on distance from STI. In practice, carrier mobility

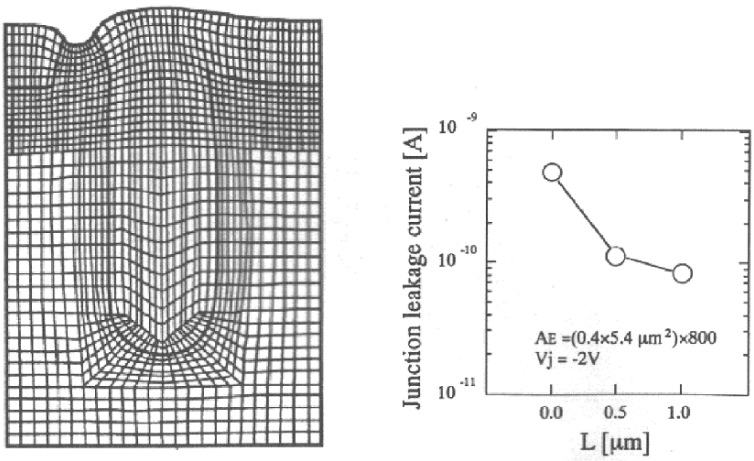


Figure 1. Displaced stress simulated mesh (left hand) and junction leakage current as a function of distance between DTI and active (right hand).

difference poses little problem for logic circuitry provided characterization for each gate (inverter, NAND, etc.) is perfect; however, it poses serious problems for analog circuitry including RF since MOSFET with multi-finger structure is often used in such circuitry. For multi-finger MOSFET, many gate fingers (channels) are available in a MOSFET and the edge of gate finger (channel) is influenced by mobility differences and the inner gate fingers (channels) are not influenced by them. Therefore, the transistor model of edge channels and that of inner channels differ. The ratio of stressed gate can be expressed as Eq. (1).

$$R_{\mathrm{STRESS}} = \frac{2nW_f}{W_g} = \frac{2nW_f}{W_f \cdot M_g} = \frac{2n}{M_g} \tag{1}$$

where, W_f is gate width for finger, M_g is number of gate fingers, W_g is total gate width, $W_g = W_f \times M_g$, and n is number of stressed gate for each side. Therefore, when number of gate finger is large, stressed gate ratio is small in other words, when gate finger width is large, stressed gate ratio is large in the case of constant total gate width.

Figure 2 shows NMOSFET transconductance dependence on gate finger width for 90 nm process with $L_g = 70\,\mathrm{nm}$, and $0.13\,\mu\mathrm{m}$ process with $L_g = 0.11\,\mu\mathrm{m}$ in the case that total gate width $= 100\,\mu\mathrm{m}$. As gate finger width is larger, STI stress appears.

To prevent this phenomenon, multi-finger MOSFET with dummy gate in both edges will be sufficient as shown in Figure 3. The upper figure of Figure 3

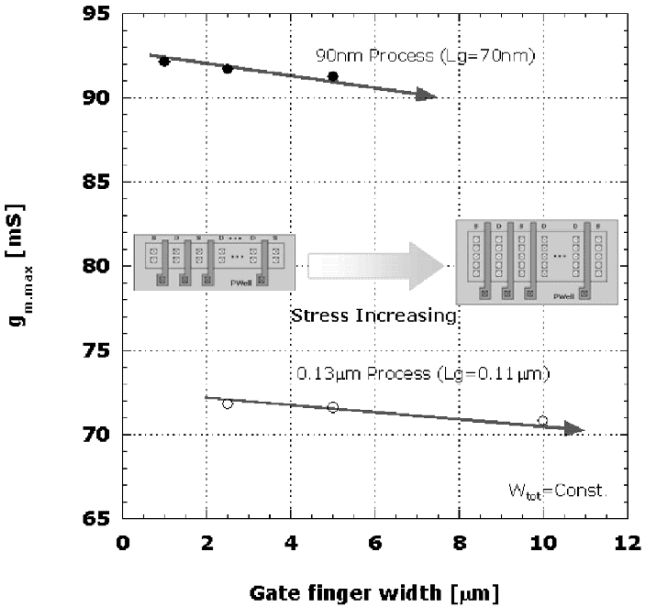


Figure 2. Transconductance degradation as a function of gate finger width.

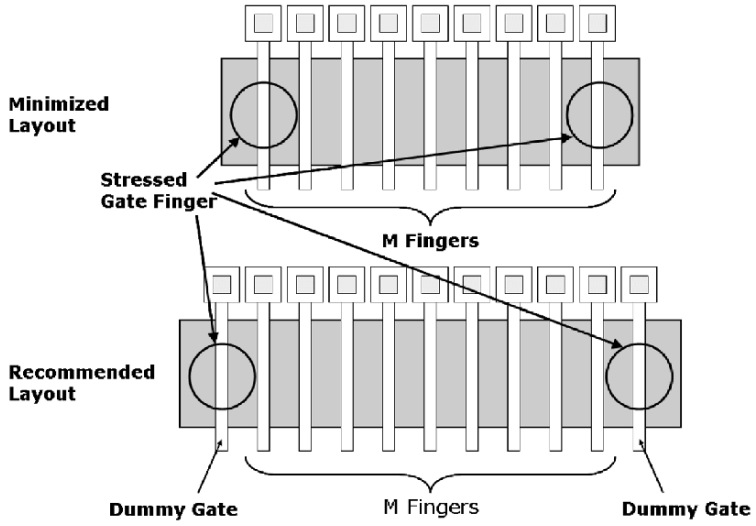


Figure 3. Stressed layout (upper) and stress-free layout (lower).

shows multi-finger MOSFET without dummy gate. This MOSFET layout makes it possible to minimize the layout but accuracy of transistor current is poor due to stressed channel. On the other hand, the lower figure of Figure 3 shows multi-finger MOSFET with dummy gate. This MOSFET layout makes it possible to obtain accurate transistor current but the layout is larger than in the case depicted in the upper figure. Since MOSFET with dummy gate may be necessary for analog circuit, the layout depicted in the lower figure is preferable for analog circuit.

2.3. Current Mirror Circuit Characteristics

The influence of STI stress was studied by using current mirror circuitry. Usually, a current mirror circuit consists of a pair of MOSFETs, one having a small number of gate fingers and the other having a large number of gate fingers and their mirror ratio is determined by only the ratio of the numbers of gate fingers. The mirror ratio is always constant except in the lower early voltage region. However, when STI stress exists, the mirror ratio is different from ideal gate finger ratio and it may depend on the difference of R_{STRESS} as represented by Eq. (1). The calculated mirror ratio as a function of R_{STRESS} is shown in Figure 4.

The designed mirror ratio of this circuitry is ten but when the stressed gate ratio increases, the accuracy of mirror ratio of current mirror circuit degrades.

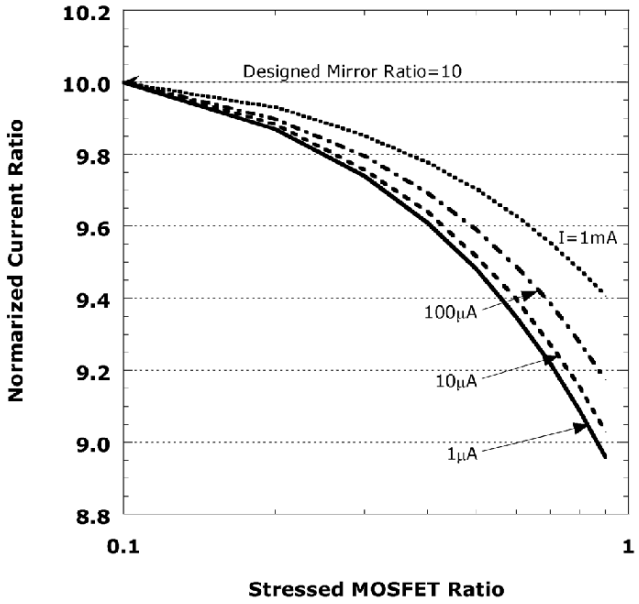


Figure 4. Mirror ratio as a function of stressed finger ratio (R_{STRESS}).

There may be a 10% difference from the designed mirror ratio in the case that R_{STRESS} is one. This difference is significantly large for analog circuit.

2.4. Summary

STI stress induced mobility degradation is on one of the most serious issues for RF circuit design and especially so in regard to transistor matching requirements with multi-finger MOSFET. Designers have to implement a dummy-gate structure or a model parameter set to cover each channel's parameter so as to prevent inaccuracy.

3. Parasitic Network Model for MOSFET

The available SPICE model of CMOS does not include parasitic components perfectly such as gate resistance, well resistance, substrate resistance, and capacitance between well and substrate. It is well known that the RF characteristic of MOSFET is strongly influenced by these parasitic components, and the influences of these components are investigated in some reports [20–24]. New types of SPICE model such as BSIM4 and/or EKV3 include the substrate network; however, the parameter values of components are set for individual

transistors even if only L_g, W_g, and M_g change. This is difficult and complicates the work of circuit designers, because many circuit designers are unfamiliar with process technology and they use a lot of MOSFETs in their circuit designs.

In this section, the scalable model of parasitic components for MOSFET is described. Each parasitic component's value can be calculated using only three basic parameters, L_g, W_f, and M_g, and the model adaptable to transistors of any size [25].

3.1. Equations for Parasitic Components

The equation for parasitic components was determined by equivalent circuit as shown in Figure 5 and target layout of MOSFET as shown in Figure 6.

The core transistor model is the normal BISM3v3 model [26] without source/drain junction capacitance (set C_j and C_{jsw} equal zero) and gate-bulk capacitance (also set C_{gbo} equal zero). The source/drain junction capacitance, gate-bulk capacitance, gate resistance, substrate resistance underlying source/drain junction, substrate resistance underlying gate-bulk capacitance, and parasitic inductance of each terminal were added to the intrinsic BSIM3v3 in this model.

Multi-finger MOSFET is commonly used in RF application to improve parasitic effects as shown in Figure 6. The present work focuses on a MOSFET that has this structure. In the case of a multi finger-MOSEFT, all finger structures are the same and the structure is repeated except at the edge part of transistors. It means some parameters, such as the distance between center of gate and back

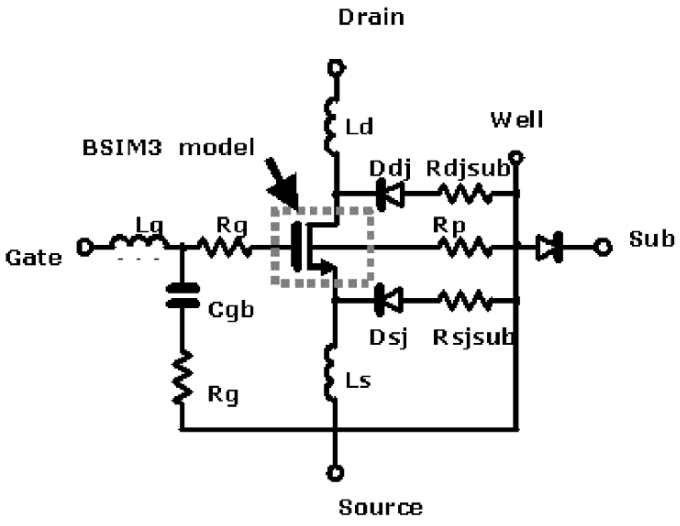


Figure 5. Equivalent circuit MOSFET for RF.

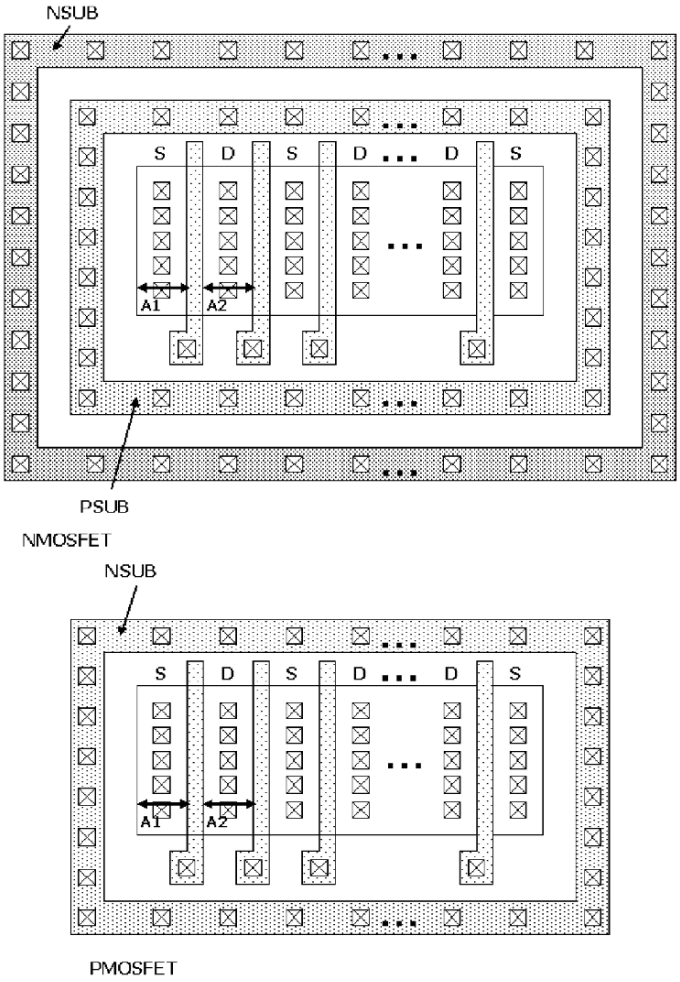


Figure 6. Target plain view of NMOS and PMOS.

gate contact and that between center of gate and substrate contact, are also the same in each fingers. Hence, the finger structure MOSFET was divided into intrinsic unit transistors for simple calculation.

3.1.1. Area and perimeter of source/drain diffusion

The source/drain diffusion of multi-finger MOSFET is common with that of the next transistor as shown in Figure 6. Thus, the number of source/drain diffusions is almost halved in this structure compared to that in a one-finger transistor. The calculated value of numbers of source/drain diffusions is shown

in Table 1, where M_s is number of source diffusion, M_d is number of drain diffusion, and M_g is number of gate fingers. In the case of an even number of gate multiples, it is possible to treat two structures of source/drain diffusion order, SDS or DSD, and the number of source/drain diffusions is different in each case. One case of the edge diffusion is source (SDS), and the other case of the edge diffusion is drain (DSD). On the other hand in the case of an odd number of gate fingers, the number of source diffusions and that of drain diffusions are the same.

The calculated equations of the area and perimeter of source/drain diffusion are shown in Table 2.

3.1.2. Gate resistance

Gate resistance, R_g, is expressed as Eq. (2).

$$R_g = R_{sg} \times \frac{(W_f + X_0)}{3 \times L_g \times M_g} + \frac{R_{cg}}{M_{cg}} \times \frac{1}{M_g} \tag{2}$$

where R_{sg} is sheet resistance of gate polysilicon, R_{cg} is contact resistance of gate polysilicon, M_{cg} is number of gate polysilicon, W_f is gate finger width, and X_o is distance between active area edge and gate polysilicon contact as shown in Figure 7 for the NMOS case. Although the layout for PMOS is not indicated here, it is almost the same as Figure 7.

3.1.3. Back gate resistance

The back gate resistance, R_n (for PMOS) and R_p (for NMOS), is dependent on length of current flow, and its length is similar to the distance between

Table 1. Number of source and drain diffusions.

# of Gate fingers	Types	M_s	M_d
EVEN	SDS	$M_g/2 + 1$	$M_g/2$
	DSD	$M_g/2$	$M_g/2 + 1$
ODD	—	$(M_g + 1)/2$	$(M_g + 1)/2$

Table 2. Area and perimeter of source and drain diffusions.

# of Gate fingers	Types	AS	PS	AD	PD
EVEN	SDS	$(M_g/2 - 1)A_2W_f + 2A_1W_f$	$(M_g - 2)(A_2 + W_f) + 4(A_1 + W_f)$	$M_gA_2W_f/2$	$M_g(A_2 + W_f)/2$
	DSD	$M_gA_2W_f/2$	$M_g(A_2 + W_f)/2$	$(M_g/2 - 1)A_2W_f + 2A_1W_f$	$(M_g - 2)(A_2 + W_f) + 4(A_1 + W_f)$
ODD	—	$(M_g - 1)A_2W_f/2 + A_1W_f$	$2(M_g - 1)(A_2 + W_f) + 2(A_1 + W_f)$	$(M_g - 1)A_2W_f/2 + A_1W_f$	$2(M_g - 1)(A_2 + W_f) + 2(A_1 + W_f)$

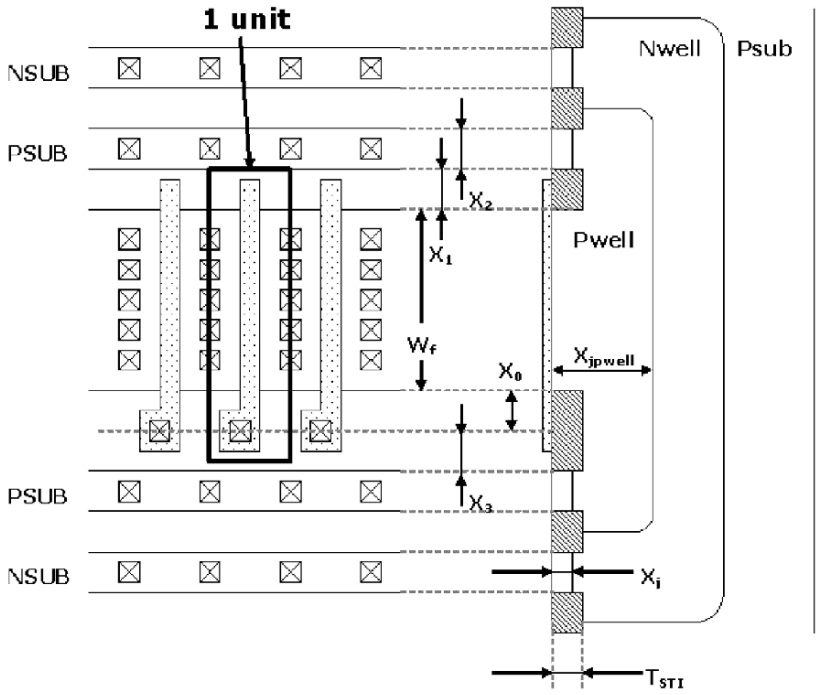


Figure 7. Details of layout view and cross section view of NMOSFET.

transistor active area and substrate contacts. To obtain accurate values of resistance, it is necessary to consider details of device structure as shown in Figure 7. Total resistance can be expressed as Eqs. (3) and (4) for NMOS and PMOS.

$$R_p = R_{spw} \times \frac{X_{\text{jpwell}}}{M_g} \times \left(\frac{X_{\text{jpwell}} + T_{\text{STI}}}{2W_f L_g} + \frac{\frac{W_f}{2} + X_1 + \frac{X_2}{2}}{(X_{\text{jpwell}} - T_{\text{STI}})^2} + \frac{X_{\text{jpwell}} - X_j}{2X_2^2} \right) \tag{3}$$

$$R_n = R_{snw} \times \frac{X_{\text{jnwell}}}{M_g} \times \left(\frac{X_{\text{jnwell}} + T_{\text{STI}}}{2W_f L_g} + \frac{\frac{W_f}{2} + X_1 + \frac{X_2}{2}}{(X_{\text{jnwell}} - T_{\text{STI}})^2} + \frac{X_{\text{jnwell}} - X_j}{2X_2^2} \right) \tag{4}$$

where $R_{spw}, R_{snw}, X_{\text{jpwell}}, X_{\text{jnwell}}, T_{\text{STI}}, X_1, X_2, X_j$ are sheet resistance of p-well, that of n-well, junction depth of p-well, that of n-well, thickness of STI, distance between gate and well contact, width of well contact, and junction depth of well contact, respectively. The first term of Eqs. (3) and (4) is resistance

of gate surface to half depth of well, the second term is resistance of gate to well contact, and the third term is resistance of half depth of well to well contact metal.

3.1.4. Well/Substrate resistance underlying source/drain diffusion

The well resistance or substrate resistance, which is underlying source/drain junction, R_{sjsub}/R_{djsub} is expressed as Eq. (5) for NMOS source, (6) for NMOS drain, (7) for PMOS source, and (8) for PMOS drain. The equation consists of components similar to those of back gate resistance.

$$R_{sjsub} = R_{spw} \times \frac{X_{\text{jpwell}}}{M_s} \times \left(\frac{X_{\text{jpwell}} - X_j}{2W_f L_{sd}} + \frac{\frac{W_f}{2} + X_1 + \frac{X_2}{2}}{(X_{\text{jpwell}} - T_{\text{STI}})^2} + \frac{X_{\text{jpwell}} - X_j}{2X_2^2} \right) \tag{5}$$

$$R_{djsub} = R_{spw} \times \frac{X_{\text{jpwell}}}{M_d} \times \left(\frac{X_{\text{jpwell}} - X_j - W_{dd}}{2W_f L_{dd}} + \frac{\frac{W_f}{2} + X_1 + \frac{X_2}{2}}{(X_{\text{jpwell}} - T_{\text{STI}})^2} + \frac{X_{\text{jpwell}} - X_j}{2X_2^2} \right) \tag{6}$$

$$R_{sjsub} = R_{snw} \times \frac{X_{\text{jnwell}}}{M_s} \times \left(\frac{X_{\text{jnwell}} - X_j}{2W_f L_{sd}} + \frac{\frac{W_f}{2} + X_1 + \frac{X_2}{2}}{(X_{\text{jnwell}} - T_{\text{STI}})^2} + \frac{X_{\text{jnwell}} - X_j}{2X_2^2} \right) \tag{7}$$

$$R_{djsub} = R_{snw} \times \frac{X_{\text{jnwell}}}{M_d} \times \left(\frac{X_{\text{jnwell}} - X_j - W_{dd}}{2W_f L_{dd}} + \frac{\frac{W_f}{2} + X_1 + \frac{X_2}{2}}{(X_{\text{jnwell}} - T_{\text{STI}})^2} + \frac{X_{\text{jnwell}} - X_j}{2X_2^2} \right) \tag{8}$$

where M_s is number of source diffusion and M_d is number of drain diffusion. Also $L_{sd} = A_s/W_f M_s$ is average value of length of source extension, $L_{dd} = A_d/W_f M_d$ is that of drain extension, and W_{dd} is depletion layer width of drain junction. Of course, W_{dd} depends on drain bias, but it is very complicated to calculate its value for each bias point. Therefore, in this work a typical biased depletion layer width was chosen (e.g. $V_{ds} = V_{gs}$).

3.1.5. Well/Substrate resistance underlying gate extension

Well/substrate resistance underlying gate extension, R_{gb}, is expressed as Eqs. (9) and (10). Equation (9) shows NMOS well/substrate resistance

underlying gate extension and Eq. (10) shows that of PMOS. The calculation methodology is also almost the same as that for back gate resistance.

$$R_{gbn} = R_{spw} \times \frac{X_{\text{jpwell}}}{M_g} \times \left(\frac{X_{\text{jpwell}} - X_j}{2A_{gf}} + \frac{X_3 + \frac{X_2}{2}}{(X_{\text{jpwell}} - T_{\text{STI}})^2} + \frac{X_{\text{jpwell}} - X_j}{2X_2^2} \right) \tag{9}$$

$$R_{gbp} = R_{snw} \times \frac{X_{\text{jnwell}}}{M_g} \times \left(\frac{X_{\text{jnwell}} - X_j}{2A_{gf}} + \frac{X_3 + \frac{X_2}{2}}{(X_{\text{jnwell}} - T_{\text{STI}})^2} + \frac{X_{\text{jnwell}} - X_j}{2X_2^2} \right) \tag{10}$$

where X_3 is distance between well/substrate contact and center of gate extension as shown in Figure 7.

3.1.6. Parasitic inductance of each terminal

Parasitic inductance of each terminal (L_s for source, L_d for drain, and L_g for gate) originates from its wire inductance. Thus, number of gate fingers is the dominant factor. The equation originated from a simple estimation of wire inductance [27] and it was optimized for adoption for some empirical results as shown in Eq. (11).

$$L_s = L_d = L_g = 1.2M_g + 18.7[\text{pH}] \tag{11}$$

All parasitic component values were calculated by Eqs. (1) to (11).

3.2. Model Confirmation

The model accuracy was confirmed by comparing between s-parameter measurement results and simulation results. MOSFET's s-parameter was measured by HP-8510 network analyzer with high frequency probe for on-wafer measurement. The parasitic capacitances in the measurement system such as pad parasitic capacitances and wire parasitic capacitances were de-embedded in an appropriate manner. Measured frequency was 0.2 to 20 GHz. Measured bias points of MOSFET were $|V_{ds}| = 1.0$ to 2.5 V, and $|V_{gs}| = 0.8$ to 1.5 V which was equivalently $|V_{th}| + 200$ mV to 900 mV. Measurement samples of MOSFET were 50 μm to 200 μm total gate width with 5 μm gate finger width, and 0.25 μm to 0.5 μm gate lengths. L_{g^-}, W_{g^-}, V_{gs^-}, and V_{ds^-} dependence were measured to compare with simulated data using this model for NMOS.

Figure 8 shows confirmation results of geometry dependence of NMOS, and Figure 9 shows that of bias dependence of NMOS, where (a) for s_{11}, (b) for s_{21}, (c) for s_{12}, and (d) for s_{22}.

Input refractions coefficient, s_{11}, differed little for different gate lengths in both simulation and measurement. Forward gain, s_{21}, shows a good agreement between simulation and measurement. Reverse gain, s_{12}, also shows good agreement between simulation and measurement, but in regions of over 10 GHz agreement is relatively poor. Output refraction coefficient, s_{22}, shows relatively poor agreement but simulated data was acceptable throughout the entire frequency range. These results indicated that this scalable parasitic model is suitable for expressing RF characteristics.

3.3. Parasitic Network Influence on RF Circuit

The influence of parasitic network on the accuracy of RF circuit simulation was confirmed. In the case of LNA, parasitic network influence on the noise figure is basically clear since the increases in noise of MOSFET results in a corresponding increase in the noise figure.

On the other hand, in the case of voltage-controlled oscillator (VCO), the parasitic network influence on the phase noise was less direct. The phase noise of VCO is influenced by several sorts of noise, including thermal noise of resonator, flicker noise of MOSFET, and also thermal noise of parasitic components of MOSFET. In this section, the phase noise differences among simulation results with substrate network, that without substrate network and measurement results are compared.

The phase noise of voltage-controlled oscillator is expressed as Eq. (12) [28].

$$L(\omega_m) = \frac{kT \cdot R_{\text{eff}}(1 + F_{GC})}{\frac{V_{\text{osc}}^2}{2}} \left(1 + \frac{\omega_{\text{osc}}}{\omega_m}\right)^2 \tag{12}$$

where, $L(\omega_m)$ is phase noise at certain hertz offset frequency ω_m from carrier, V_{osc} is oscillation amplitude of VCO, ω_{osc} is oscillation frequency, and F is noise parameter. Although the definitions of almost all the parameters in this equation are clear, a part of F is still unclear. In ideal VCO, F_{GC} consists of the resonator noise source and the gain-cell noise source.

Noise sources of resonator are fundamental phase noise contents of integrated VCO, and several papers have reported on them [27–29]. The effective resistance of resonator is expressed as Eq. (13).

$$R_{\text{eff}} = R_l + R_{vc} + \frac{1}{R_p(\omega_{\text{osc}} C_{\text{tot}})^2} \tag{13}$$

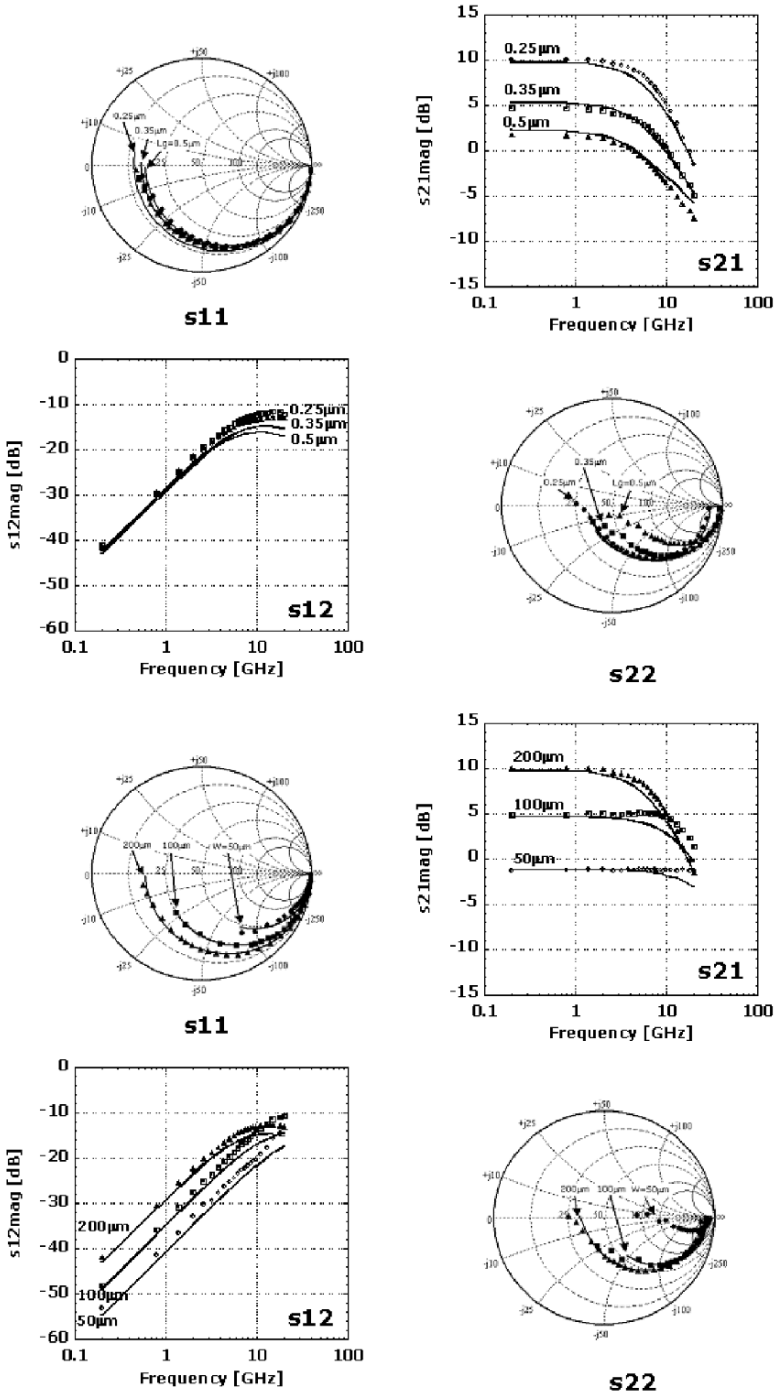


Figure 8. Measurement result and simulated one as a function of L_g and W_g.

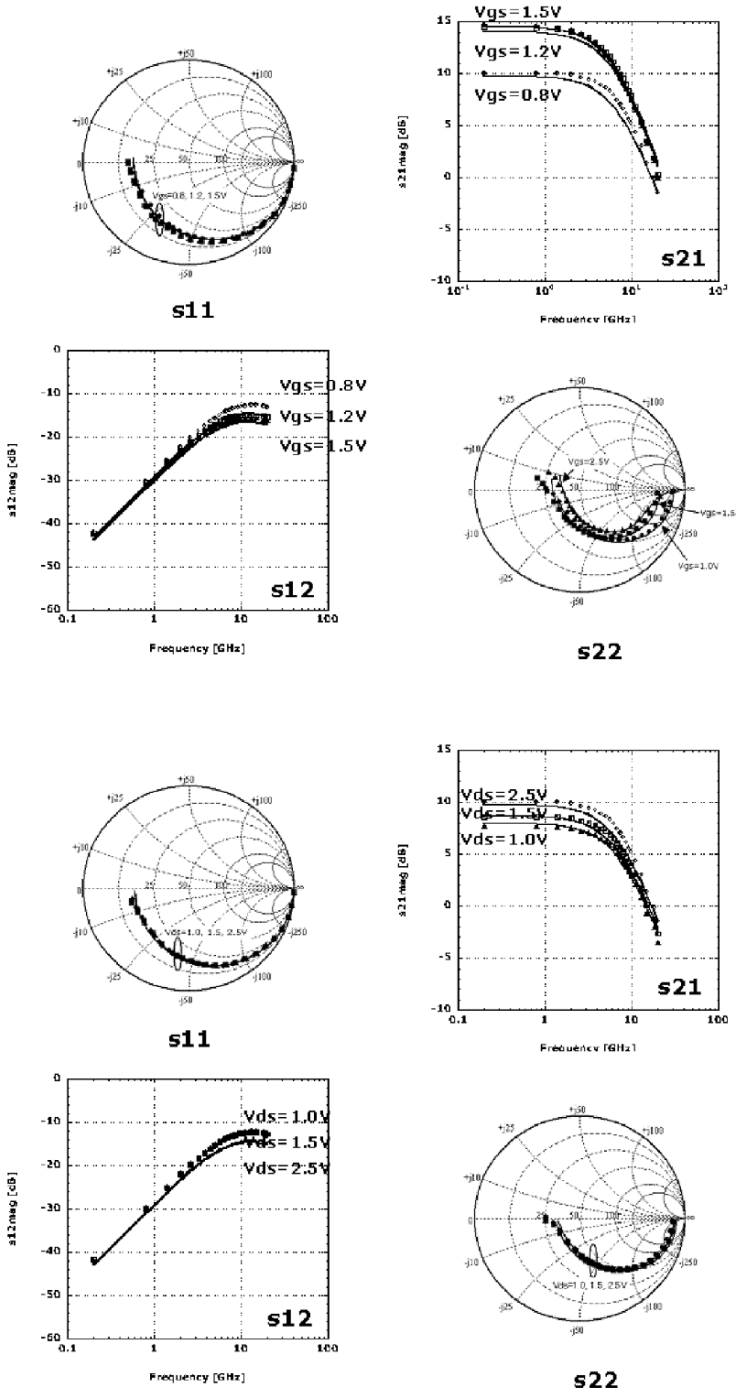


Figure 9. Measurement result and simulated one as a function of DC-bias.

where, R_l is parasitic resistance of inductor, R_{cv} is parasitic resistance of varactor, R_p is parallel resistance of resonator, and C_{tot} is total capacitance of resonator including varactor capacitance and any parasitic capacitance, where, G_m of MOS-VCO gain-cell and noise equation of MOS gain-cell is expressed as Eq. (15).

$$\overline{di}^2_{M,d} = 4kT\left(\sum \gamma g_{d0} + \sum R_i g_{d0}^2\right)\Delta f \tag{15}$$

This section focuses on parasitic resistance, hence right hand of Eq. (15) is significant. The correct noise contribution factor of MOS-VCO gain-cell from the viewpoint of parasitic resistance is expressed as Eq. (16).

$$\alpha_{\text{Noise}} = \sum R_i g_{d0} \tag{16}$$

The total noise equation of phase noise of MOS-VCO is rewritten as Eq. (17).

$$L(\omega_m) = \frac{kT \cdot R_{\text{eff}}(1 + \sum R_i g_{d0})}{\frac{V_{\text{osc}}^2}{2}}\left(1 + \frac{\omega_{\text{osc}}}{\omega_m}\right)^2 \tag{17}$$

Phase noise was calculated using Eqs. (11), (13), and (15). Figure 10 shows simulation results of MOS-VCO phase noise at 3 MHz offset from carrier with

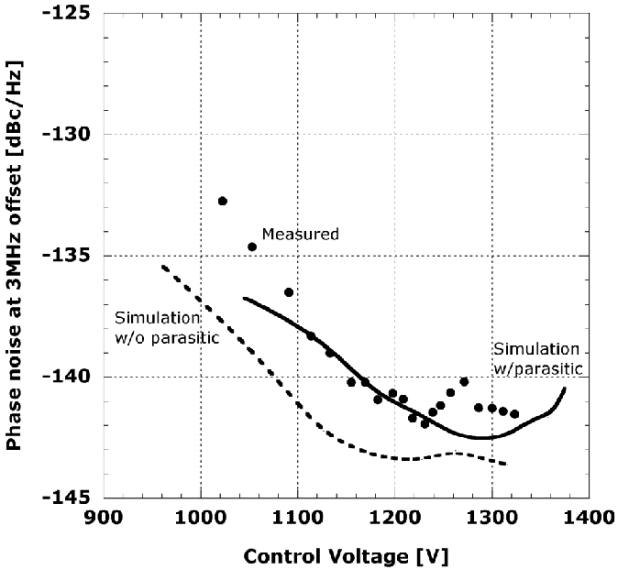


Figure 10. The difference of phase noise among measured, simulated with parasitic, and simulated without parasitic.

parasitic network model and that without parasitic network model, and measurement data. Measurement data shows good agreement with simulation data with parasitic network. On the other hand, simulation data without parasitic network shows disagreement with measurement data. The importance of the parasitic network is clarified.

3.4. Summary

The parasitic components model is very important for RF circuit design and its influence is significant as shown in this section, not only for small signal parameter accuracy but also for large signal circuit such as VCO. Introduction of a scalable parasitic model will be necessary for modern circuit design.

4. Channel Noise

4.1. Channel Noise of Small Geometry MOSFET

A simplified MOSFET equivalent circuit with noise sources is shown in Figure 11.

Noise sources of MOSFET consist of gate resistance noise, source resistance noise, drain resistance noise, flicker noise, body resistance noise, and channel thermal noise. The five noises other than channel thermal noise can be expressed

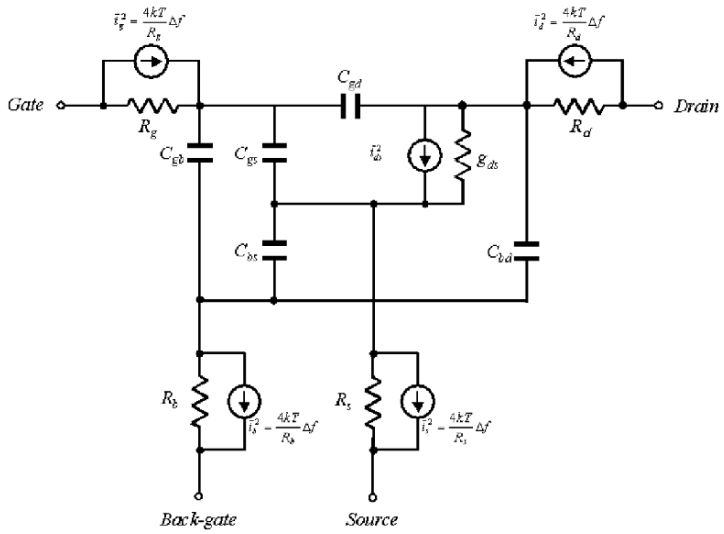


Figure 11. Noise equivalent circuit of MOSFET.

as the following equations.

$$\bar{i}_g^2 = \frac{4kT}{R_g}\Delta f \tag{18}$$

$$\bar{i}_s^2 = \frac{4kT}{R_s}\Delta f \tag{19}$$

$$\bar{i}_d^2 = \frac{4kT}{R_d}\Delta f \tag{20}$$

$$\bar{i}_{1/f}^2 = \frac{K_F I_{ds}^{A_F}}{f \cdot C_{OX} \cdot W_g \cdot L_g}\Delta f \tag{21}$$

$$\bar{i}_b^2 = \frac{4kT}{R_b}\Delta f \tag{22}$$

where, k is Boltzman's constant, T is absolute temperature, R_g is gate resistance, R_s is source resistance, R_d is drain resistance, K_F is flicker noise coefficient, A_F is noise exponential coefficient, C_{ox} is gate insulator capacitance, W_g is gate width, L_g is gate length, and R_b is total body resistance.

The channel thermal noise of recent small geometry MOSET consists of two regions as shown in Figure 12. One is gradual electron velocity region and the other is velocity saturation region. Both regions are divided at pinch-off point. Channel length of gradual electron velocity region is L_{elec} and that of velocity saturation region is ΔL in Figure 12. An applied drain to source voltage of gradual electron velocity region is $V_{d\text{sat}}$ in total drain to source voltage, V_{ds}.

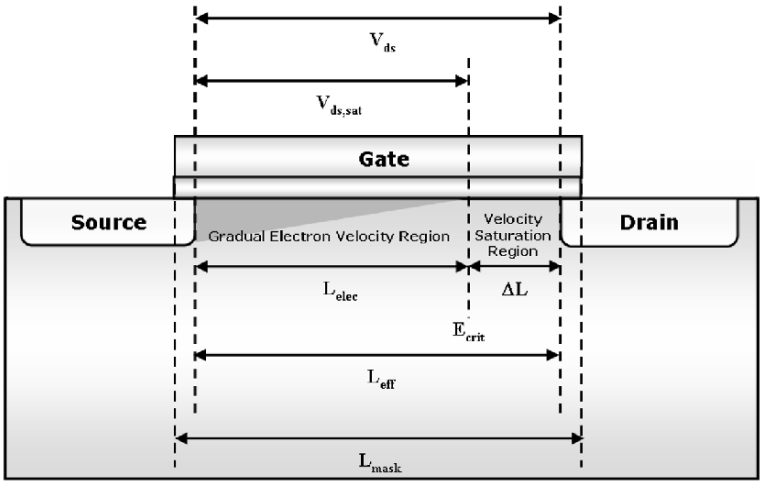


Figure 12. Cross-section view of MOSFET.

The channel thermal noise of gradual electron velocity region can be expressed as Eq. (23) [30].

$$\overline{i_{ch}^2} = 4kT\left[\gamma\frac{W_g}{L_g}\mu C_{OX}\left(V_{gs} - V_{th}\right)\right] = 4kT\gamma g_{d0}$$
$$\gamma = \frac{2}{3}\times\frac{1+\eta+\eta^2}{1+\eta} \tag{23}$$

where γ is channel thermal noise coefficient, μ is mobility of carrier, V_{gs} is gate to source voltage, V_{th} is threshold voltage of MOSFET, η is parameter of noise, and g_{d0} is zero biased drain to source conductance. When MOSFET works in linear region, η is unity and when MOSFET works in saturation region, η is zero. Hence γ is 1 and 2/3 in the case of linear region and saturation region of MOSFET, respectively. This channel thermal noise coefficient is the same as the classical one.

The channel thermal noise in velocity saturation region can be expressed as Eq. (24) [9].

$$\overline{i_{ch,vs}^2} = \delta\frac{4kT}{L_g^2}\cdot\frac{I_{ds}}{E_{\text{crit}}}\frac{1}{\alpha}\sinh\left(\alpha\Delta L\right) \tag{24}$$

where ΔL is channel length of velocity saturation region as shown in Eq. (25), E_{crit} is critical electric field along channel, and δ is a fitting parameter. The channel thermal noise of this region is defined by hot electron regime.

$$\Delta L = \frac{1}{\alpha}\ln\left[\frac{\alpha\left(V_{ds} - V_{d\text{sat}}\right) + E_D}{E_{\text{crit}}}\right] \tag{25}$$

$$E_D = E_{\text{crit}}\sqrt{1+\left[\frac{\alpha\left(V_{ds} - V_{d\text{sat}}\right)}{E_{\text{crit}}}\right]^2} \tag{26}$$

$$\alpha = \lambda\sqrt{\frac{3}{2}\cdot\frac{C_{OX}}{x_j\cdot\varepsilon_{Si}\cdot\varepsilon_0}} \tag{27}$$

where x_j is the junction depth of source/drain and λ is a fitting parameter of channel length modulation. To define enhancement of the channel thermal noise due to scaling, we measured MOSFET's noise and determined channel thermal noise with different gate length [31].

4.2. Channel Noise Measurement and Characterization

The noise figure of device was measured at frequencies of 1 to 6 GHz. The measurement configuration is shown in Figure 13. The device was measured

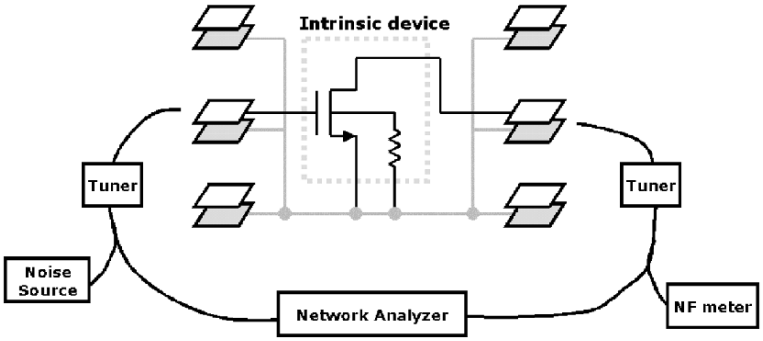


Figure 13. Set-up configuration for on-wafer RF measurement.

by common source and input and output terminals were connected to shielded GSG pads which can eliminate body noise. The parasitic capacitance, parasitic inductance and parasitic resistance were de-embedded by replica pads and wired measurement pattern. The input and output impedances were measured by vector network analyzer (NWA) and tuned by tuner of respective terminals. The noise was measured by NF meter.

Measured geometry of NMOS, L_g/W_g, were 40 nm/100 μm, 60 nm/100 μm, and 70 nm/100 μm with 90 nm process, 110 nm/100 μm with 130 nm process, and 140 nm/100 μm with 180 nm process. The gate width of each MOSFET consisted of 20 × 5 μm finger structure. This means gate width of MOSFET was very large, and therefore, parasitic resistance of source terminal and parasitic resistance of drain terminal can be negligible. Measurement conditions were $V_{ds} = 1$ V and several V_{gs}.

The $NF_{\min}$ was carried out by equivalent noise circle in smith chart. The data were determined by measurement data of several input and output matching conditions. Measured $NF_{\min}$ is dependent on drain current as shown in Figure 14. Due to scaling down of MOSFET gate length, $NF_{\min}$ decreased by 70 nm. However, $NF_{\min}$ did not improve below 70 nm gate length. It is thought that MOSFET noise originating from either gate resistance or channel resistance increases due to scaling down.

In order to extract channel thermal noise, we measured 50 Ω termination noise figures, NF50. The frequency response of NF50 is shown in Figure 15. Generally, noise figure of MOSFET shows frequency response. In the case of low frequency, the noise increases due to influence of flicker noise. On the other hand, in the case of high frequency, the noise increases, too, as a result of gain degradation due to high frequency. In order to obtain channel thermal noise correctly, it is necessary to use mid-frequency range. In Figure 15, in the frequency range above 4 GHz, NF50 increases for almost all MOSFETs regardless

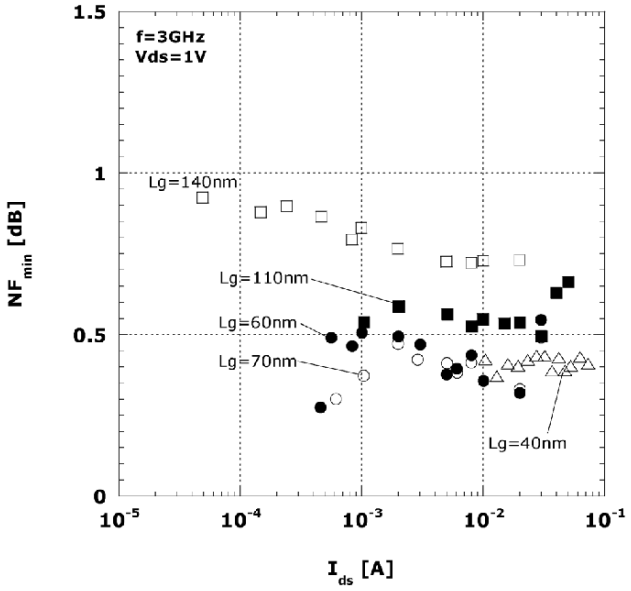


Figure 14. Drain current dependence on measured $NF_{\min}$.

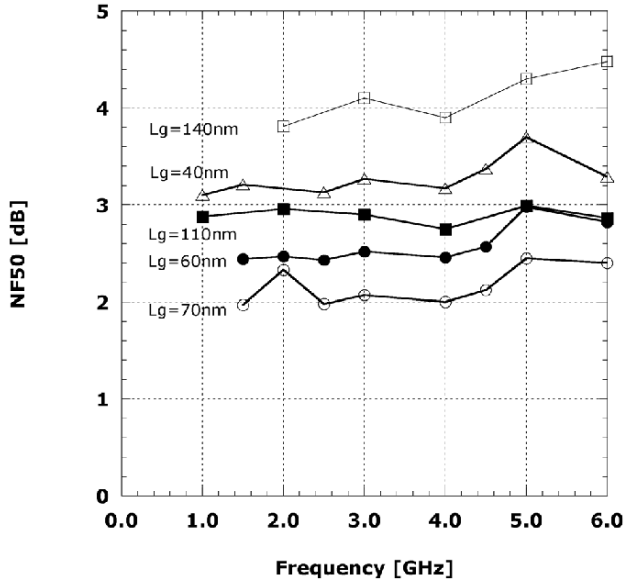


Figure 15. The operating frequency dependence on NF50.

of gate length. It was supposed that there were two reasons for these phenomena. One was MOSFET's gain degradation due to high frequency operation and another was measurement instability. Therefore, frequency range of 1 to 4 GHz was chosen to extract channel thermal noise to obtain correct channel noise performance.

Measured NF50 data for this frequency range indicated the virtual elimination of body resistance noise, source resistance noise, drain resistance noise, and flicker noise. Hence it only contains channel thermal noise and gate resistance noise. The gate resistance can be calculated by Eq. (2). The channel thermal noise was extracted by subtracting gate resistance noise from total noise which carried out NF50. The extracted channel thermal noise as a function of gate overdrive voltage, V_{gs}-V_{th}, is shown in Figure 16.

The channel thermal noise was approximately $3.0 \times 10^{-21}, 2.5 \times 10^{-21}, 2.2 \times 10^{-21}, 1.8 \times 10^{-21}, 1.6 \times 10^{-21}$, for 40 nm, 60 nm, 70 nm, 110 nm, and 140 nm gate length MOSFET at 0.3 V gate overdrive voltage. Indeed, the results indicate the channel thermal noise increased due to scaling down, and the channel thermal noise of 40 nm gate length NMOS was over two times larger than that of 140 nm gate length.

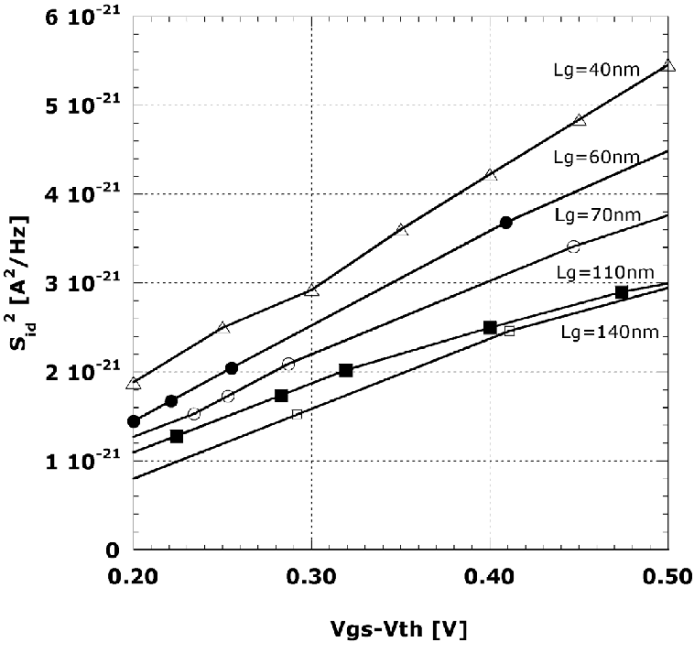


Figure 16. Gate overdrive dependence of noise current.

The total channel thermal noise in saturation region can be express as Eq. (28) using Eqs. (23) and (24).

$$\begin{aligned}\overline{i_{ch,vs}^2} &= 4kT\gamma g_{d0} + \delta\frac{4kT}{L_g^2}\cdot\frac{I_{ds}}{E_{\text{crit}}}\frac{1}{\alpha}\sinh(\alpha\Delta L)\\ &= 4kTg_{d0}\left(\gamma + \delta\frac{4kT}{L_g^2}\cdot\frac{I_{ds}}{E_{\text{crit}}}\frac{1}{\alpha}\sinh(\alpha\Delta L)\frac{1}{g_{d0}}\right)\\ &= 4kTg_{d0}\gamma_{em}\end{aligned} \tag{28}$$

where, γ_{em} is empirical notation of noise coefficient, which covers from long channel to sub 0.1 μm channel. γ_{em} can be rewritten as Eq. (29).

$$\gamma_{em} = \frac{2}{3} + \delta\frac{I_{ds}}{L_g^2\cdot E_{\text{crit}}\cdot\alpha}\sinh(\alpha\Delta L)\frac{1}{g_{do}} \tag{29}$$

The first term of Eq. (28) indicates classical channel thermal noise in [30] and the second term indicates empirical equation similar to [10]. The measured data and calculated results were compared at around $g_{m,\max}$ point for each NMOS. The calculation results of γ_{em} by Eq. (29), measured data of this work, and some published data with similar bias condition are shown in Figure 17. Figure 17 shows quite good agreement from few μm gate lengths to sub 0.1 μm gate length. The calculated channel thermal noise coefficient, γ, were 3.5, 2.2,

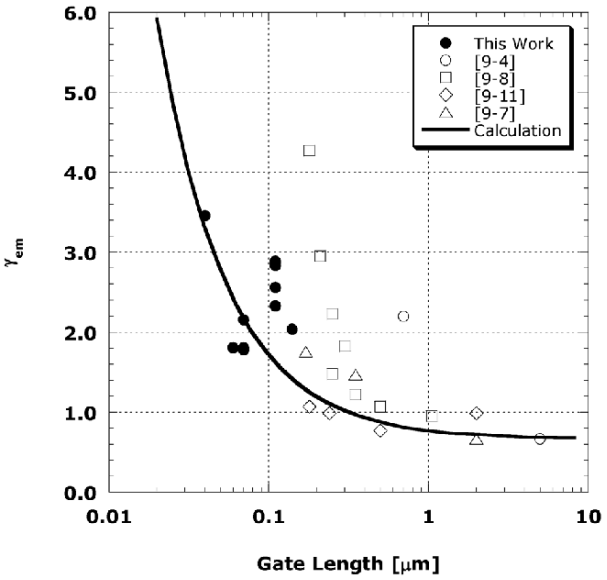


Figure 17. Channel thermal noise coefficient, γ, as a function of gate length of this work and published data. The solid line was calculated by Eq. (29).

2.0, 1.6, and 1.5 for 40 nm, 60 nm, 70 nm, 110 nm, and 140 nm gate length NMOS, respectively. The γ of 40 nm gate length NMOS was approximately five times larger than that of long channel NMOS. The calculation curve almost fit the measurement data of this work and also almost fit the published data. In this fitting curve, fitting parameters were set as $\delta \sim 10$ for Eq. (28) and $\lambda = 0.65$ to 0.95 for Eq. (27).

4.3. Influence for Phase Noise Calculation of VCO

Phase noise of integrated VCO without flicker noise contribution is expressed as Eq. (12) and noise equation of MOS gain-cell is expressed as Eq. (15).

Therefore, correct noise contribution factor of MOS-VCO gain-cell is expressed as Eq. (30) [29].

$$F_{GC} = \alpha_{\text{Noise}} A = \gamma A \tag{30}$$

The total noise equation of phase noise of MOS-VCO is rewritten as Eq. (31).

$$L(\omega_m) = \frac{kT \cdot R_{\text{eff}}(1 + \gamma A)}{\frac{V_{\text{osc}}^2}{2}} \left(1 + \frac{\omega_{\text{osc}}}{\omega_m}\right)^2 \tag{31}$$

To confirm the calculation accuracy of Eq. (12), Figure 18 shows a comparison of calculation results obtained by Eq. (13) and measured phase noise data of several VCOs using 0.25 μm to 0.5 μm gate lengths MOSFET. Compared offset frequency from carrier was 1 MHz and compared control voltage was over 1.0 V to avoid any other component influences such as flicker noise contribution and current noise contribution from current source in this work. The closed circle in Figure 18 indicates calculation using optimum γ value which was extracted by Eq. (13) and the open circle indicates calculation using constant γ as 2/3. Indeed, this figure shows the optimum γ is in better agreement with measured data than is the constant γ value of 2/3. The accuracy using optimum γ for analytical expression of VCO was within $\pm 2\,dB$ on average but that using constant γ was over $\pm 2\,dB$.

4.4. Summary

The channel thermal noise coefficient, γ, was extracted by a high frequency measurement method. It was correctly extracted, eliminating any other noise sources such as source resistance, drain resistance, flicker noise, body resistance, and gate resistance. Indeed, in the case of small gate length MOSFET,

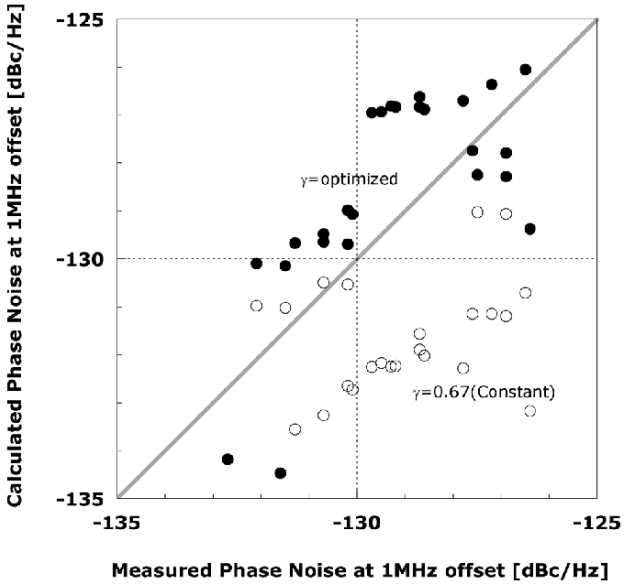


Figure 18. Comparison of measure phase noise and calculated phase noise of several fully integrated MOS-VCO with optimum γ and constant γ as 2/3. Phase noise was measured and calculated at 1 MHz offset from carrier.

γ increased due to hot carrier effect. It was approximately five times larger than classical noise coefficient value, 2/3, in the case of 40 nm MOSFET. The empirical equation of channel thermal noise was in quite good agreement with measured data.

The noise coefficient enhancement influences RF circuit performance. In this work, the influence of increased phase noise of MOS-VCO was confirmed using analytical expression of VCO phase noise. The calculation accuracy using the analytical expression of phase noise of VCO and empirical noise equation of MOS-VCO was within $\pm 2\,dB$.

5. Conclusion

In this chapter, some parameters not covered by a compact model, such as STI stress, scalable parasitic components model and channel thermal noise enhancement due to scaling were described with some circuit performances.

At least, these three issues should be solved by achieving accuracy of these models, which will shrink both the cost and design period for complicated RF circuit design.

Acknowledgement

The author is grateful to Dr. S. Yoshitomi, Mr. T. Ohguro and Mr. K. Kojima of Semiconductor Company, Toshiba Corporation, for fruitful discussion regarding this work. He also wishes to thank Mr. S. Watanabe and Mr. H. Aoki of Semiconductor Company, Toshiba Corporation, for support throughout this work.

References

[1] Steyaert, M.; Janssens, J.; De Muer, B.; Borremans, M.; Itoh, N. "A 2 volt CMOS cellular transceiver front-end", *IEEE J. Solid-State Circuits*, **December 2000**, *35(12)*, 1895–1907.

[2] Itoh, N. "MOSFET modeling for RF circuit design", Presentation of MOS-AK work shop, Leuven, **September 2004**.

[3] Bianchi, R.; Bouche, G.; Roux-dit-Buisson, O. "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance", *Proc. Int. Electron Dev. Meeting*, **2002**.

[4] Abidi, A.A. "High-frequency noise measurements on FETs with small dimensions", *IEEE Trans. Electron Dev.*, **November 1986**, *ED-33*, 1801–1805.

[5] Triantis, D.P.; Birbas, A.N.; Kondis, D. "Thermal noise modeling for short-channel MOSFET's", *IEEE Trans. Electron Dev.*, **November 1996**, *43(11)*, 1950–1955.

[6] Klein, P. "An analytical thermal noise model of deep submicorn MOSFET's", *IEEE Electron Dev. Lett.*, **August 1999**, *20(8)*, 399–401.

[7] Scholten, A.J.; Tromp, H.J.; Tiemeijer, L.F.; van Langevelde, R.; Havens, R.J.; de Vreede, P.W.H.; Roes, P.F. M.; Woerlee, P.H.; Montree, A.H.; Klassen, D.B.M. "Accurate thermal noise modeling for deep-submicron CMOS", *Proc. Int. Electron Dev. Meeting*, **December 1999**, 155–158.

[8] Knoblinger, G.; Klein, P.; Baumann, U. "Thermal channel noise of quarter and subquarter micron NMOS FET's", *Proc. ICMTS*, **2000**, 95–98.

[9] Enz, C.C.; Cheng, Y. "MOS transistor modeling for RF IC design", *IEEE Trans. Solid-State Circuits*, **February 2000**, *35(2)*, 186–201.

[10] Knoblinger, G.; Klein, P.; Tiebout, M. "A new model for thermal channel noise of deep-submicron MOSFETS and its application in RF-CMOS design", *IEEE J. Solid-State Circuits*, **May 2001**, *36(5)*, 831–837.

[11] Scholten, A.J.; Tiemeijer, L.F.; van Langevelde, R.; Havens, R.J.; Venezia, V.C.; Zagers-van Duijnhoven, A.T.A.; Neinhus, B.; Jungemann, C.; Klassen, D.B.M. "Compact modeling of drain and gate current for RF CMOS", *Proc. Int. Elecron Dev. Meeting*, **2002**, 129–132.

[12] Scholten, A.J.; Tiemeijer, L.F.; van Langevelde, R.; Havens, R.J.; Zagers-van Duijnhoven, A.T.A.; Venezia, V.C. "Noise modeling for RF CMOS circuit simulation", *IEEE Transaction on Electron Dev.*, **March 2003**, *50(3)*, 618–632.

[13] Deen, M.J.; Chen, C.H.; Cheng, Y. "MOSFET modeling for low noise, RF circuit design", *Proc. Custom Integrated Circuit Conf.*, **2002**, 201–208.

[14] Chen, C.H.; Deen, M.J. "Channel noise modeling of deep submicron MOSFET's", *IEEE Trans. Electron Dev.*, **August 2002**, *49(8)*, 1484–1487.

[15] Asgaran, S.; Deen, M.J.; Chen, C.H. "Analytical modeling of MOSFET noise parameters for analog and RF applications", *Proc. Custom Integrated Circuit Conf.*, **2004**, 379–382.

[16] Koeppe, J.; Harjani, R. "Enhanced analytic noise model for RF CMOS design", *Proc. Custom Integrated Circuit Conf.*, **2004**, 383–386.

[17] Itoh, N.; Yoshino, C.; Matsuda, S.; Tsuboi, Y.; Inou, K.; Katsumata, Y.; Iwai, H. "Optimization of shallow and deep trench isolation structures for ultra-high-speed bipolar LSIs", *Proc. 1992 IEEE Bipolar/BiCMOS Circuits Technol. Meeting*, Minneapolis, **September 1992**, 104–107.

[18] Matsuda, S.; Itoh, N.; Yoshino, C.; suboi, Y.; Katsumata, Y.; Iwai, H. "Mechanical stress analysis of trench isolation using a two-dimensional simulation", *IEICE Trans. Electron.*, **February 1994**, *E77-C(2)*, 124–128.

[19] Bianchi, R.A.; Bouche, G.; Roux-dit-Buisson, O. "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance", *Proc. IEDM*, **2002**.

[20] Tin, S.F.; Osman, A.A.; Mayaram, K.; Hu, C. "A simple subcircuit extension of BSIM3v3 models for CMOS RF design", *IEEE J. Solid-State Circuits*, **April 2000**, *35(4)*, 612–623.

[21] Deen, M.J.; Chen, C.H. "MOSFET modeling for low noise, RF circuit design", *Proc. IEEE 2002 Custom Integrated Circuit Conf.*, **May 2002**, 201–208.

[22] Enz, C.C.; Cheng, Y. "MOS transistor modeling for RF IC design", *IEEE Trans. Solid-State Circuits*, **February 2000**, *35(2)*, pp186–199.

[23] Fujimoto, R.; Watanabe, O.; Fujii, F.; Kawakita, H.; Tanimoto, H. "High-frequency device-modeling techniques for RF-CMOS circuits", *IEICE Trans. Fundamentals*, **February 2001**, *E84-A(2)*, 520–528.

[24] Liu, W.; Gharpurey, R.; Chang, M.C.; Edogan, U.; Aggarwal, R.; Mattia, J.P. "R.F. MOSFET modeling accounting for distributed substrate and channel resistances with emphasis on the BSIM3v3 SPICE model", *IEEE IEDM Tech. Dig.*, **1997**, 309–312.

[25] Itoh, N.; Ohguro, T.; Katoh, K.; Kimijima, H.; Ishizuka, S.; Kojima, K.; Miyakawa, H. "Scalable parasitic components model of CMOS for RF circuit design", *IEICE Trans. Fundamentals*, **February 2003**, *E86-A(2)*, 288–298.

[26] BSIM3 version 3 Manual, Department of Electrical Engineering and Computer Sciences, University of California, Barkely, CA, **1996**.

[27] Lee, T. *The Design of CMOS Radio Frequency Integrated Circuits.* Cambridge University Press, **1998**, ISBN 0-521-63061-4.

[28] Craninckx, J.; Steyaert, M. "A 1.8-GHz CMOS low-phase-noise voltage-controlled oscillator", *IEEE J. Solid-State Circuits*, **December 1995**, *30(12)*, 1474–1482.

[29] Steyaert, M.; Borremans, M.; Janssens, J.; De Muer, B.; Itoh, N.; Craninckx, J.; Crols, J.; Morifuji, E.; Momose, H.S.; Sansen, W. "A single-chip CMOS transceiver front-end for DCS-1800 wireless communications", *Analog Integrated Circuits and Signal Processing*, **August 2000**, *24(2)*, 83–99.

[30] Tsividis, Y.P. "Operation and modeling of the MOS transistor", New York: McGraw-Hill, **1988**.

[31] Ko, P.K.; Muller, R.S.; Hu, C. "A unified model for the hot-electron currents in MOSFET's", *Proc. Int. Electron Dev. Meeting*, **1981**, 600–603.

Chapter 8

ON INCORPORATING PARASITIC QUANTUM EFFECTS IN CLASSICAL CIRCUIT SIMULATIONS

Frank Felgenhauer, Maik Begoin and Wolfgang Mathis
University of Hannover, Institute of Electromagnetic Theory and Microwave Technique, Appelstraÿe 9A, 30167 Hannover, Germany
E-mail: (felgenhauer, begoin, mathis)@tet.uni-hannover.de

Abstract: In this chapter, we present a discussion about the influence of parasitic quantum effects to the functionality of classical electronic circuit concepts. The discussion covers the physics and the simulation of coherent charge transport and also the way to include quantum effects in high level circuit simulators like SPICE. Electronic circuits we are talking about are scaled into a domain where the common semi-classical transport models loose more and more their validity. Therefore, we start with a review of the semi-classical semiconductor equations and their extensions to include quantum effects. Further, a derivation of the quantum transport equations for coherent electron transport is given, including a short summary of current methods to solve these equations. The Schrödinger-Poisson solver we use to calculate transport is presented in detail. At the end of the chapter we show three different circuit examples, which explicitly exhibit the influence of quantum effects to circuit functionality.

Key words: CMOS circuits; parasitic quantum effects; SPICE simulation

Introduction

Although several new device concepts are considered in nanotechnology during the last decade industrial applications will be dominated by CMOS technology in the near future since very complex CMOS circuits can be realized. Due to the rapid process of down-scaling of integrated semiconductor devices

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, 209–241.

the performance of a CMOS device is influenced by an increasing number of parasitic effects. Beside semi-classical parasitic effects and leakage currents such as sub-threshold currents, DIBL and GIBL [1] that have already taken into account in the sub-μm regime further parasitic effects of quantum mechanical origin must be included in device modelling. One of the well-known quantum mechanical effects is the tunneling current through thin potential barriers. For example, in 0.12 μm technology the oxide thicknesses of gates is below 3 nm and the direct tunneling current starts to increase exponentially [2, 3]. In consequence, for classical CMOS circuits we have to expect at least a dramatic increase of these parasitic currents leading to unacceptable noise levels in analog applications [4] and may even cause a failure of the circuit functionality [5]. However, there are more quantum mechanical phenomena, like charge quantization, scattering etc., which can restrict the functionality of classical circuits. In this chapter, we present our methodology to analyze nanoscaled circuits, from the physics of charge carrier transport to high-level, SPICE-like circuit simulators. We start the first section with a discussion of semi-classical semiconductor equations and their limits, when devices and structures in the mesoscopic regime are considered. Subsequent, we show a structural derivation of basic quantum transport equations in multi-layered semiconductors, since circuit designers are usually not familiar with physics of charge carrier transport in-depth. This section ends with a short summary about methods to solve the semi-classical and quantum mechanical semiconductors equations. In the following two sections we present in detail a self-consistent Schrödinger Poisson solver based on the non-equilibrium Green's function formalism (NEGF) and a Newton-Raphson algorithm. As an example, the calculation of coherent transport through isolating oxide layer, which corresponds to gate direct tunneling in MOSFET, is shown. Let us note at this point that the transport processes we consider in this chapter are all coherent since the considered nanoscaled devices and structures below 20 nm are much shorter than the coherence length. However, the algorithm we use to solve the Schrödinger equation (e.g. the NEGF formalism) can be extended to incoherent transport.

The last section is about the incorporation of the results of the quantum mechanical transport simulations into a high-level circuit simulator. We present three different circuit examples, each representing a specific circuit type and each example exhibits the influence of direct tunneling currents to the circuit functionality.

1. From Drift-Diffusion to Wavelike Behaviour

Devices in electronic circuits are connected to at least two contacts, therefore any device we are talking about is an open system with respect of charge carrier transport. In principle, transport processes in conductors and semiconductors

have to be described as many-body problems where the dynamics of particles have to be considered by an "ensemble" description instead of a single particle description or "test particle" description. This can be done for classical as well as quantum mechanical transport processes.

1.1. Semi-classical Transport

Classical transport processes are based on the classical Newton mechanics where we replace the single particle dynamics with the force $\mathbf{F} : I\!R^3 \to I\!R^3$ by the dynamics of a whole set of single particles which can be interpreted as an "ensemble" of particles. For this purpose we define a time-dependent probability distribution function $f(\mathbf{r}, \mathbf{v}, t)$ on the "state space" $(\mathbf{r}, \mathbf{v}) \in I\!R^3 \times I\!R^3$ of a one-particle system where $\mathbf{v}$ is the velocity of the particles and study their dynamics. Using $f(\mathbf{r}, \mathbf{v}, t)$ the number of particles at time t in a volume V can be calculated as

$$N(t) := \int f(\mathbf{r}, \mathbf{v}, t)\, d^3r d^3v. \tag{1}$$

The number of particles in V changes with t because some particles enter as well as leave V. At first we assume that there are no collisions between the particles. Then if a single particle is in a state $(\mathbf{r}, \mathbf{v})$ at time t it will be in the state $(\mathbf{r} + \mathbf{v}\delta t, \mathbf{v} + (\mathbf{F}/m)\delta t)$. However we have

$$f\left(\mathbf{r} + \mathbf{v}\delta t, \mathbf{v} + \frac{\mathbf{F}}{m}\delta t, t + \delta t\right) = f(\mathbf{r}, \mathbf{v}, t). \tag{2}$$

If collisions occur an additional collision term has to be taken into account

$$f\left(\mathbf{r} + \mathbf{v}\delta t, \mathbf{v} + \frac{\mathbf{F}}{m}\delta t, t + \delta t\right) = f(\mathbf{r}, \mathbf{v}, t) + \left(\frac{\partial f}{\partial t}\right)_{\text{coll}} \delta t. \tag{3}$$

If the left hand side of (3) is to be developed to the first order we obtain *Boltzmann's equation* of kinetic theory (see e.g. Huang [6])

$$\left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla_r + + \frac{\mathbf{F}}{m} \cdot \nabla_v\right) f(\mathbf{r}, \mathbf{v}, t) = \left(\frac{\partial f}{\partial t}\right)_{\text{coll}}. \tag{4}$$

The second term of the left side of Eq. (4) is called *diffusion term* whereas the third term is called *drift term*. Note that Boltzmann's equation can be interpreted only if the collision term is defined explicitly.

If we restrict ourselves to two-particle interactions and molecular chaos is assumed an explicit collision term can be derived

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = \int d\Omega \int d^3v_2 \sigma(\Omega) \|\mathbf{v}_1 - \mathbf{v}_2\| (f_2' f_1 - f_2 f_1) \tag{5}$$

where Ω is the angle between $\mathbf{v}_1 - \mathbf{v}_2$ and $\mathbf{v}_1' - \mathbf{v}_2'$, $f_1 := f(\mathbf{r}, \mathbf{v}_1, t), \ldots$ and $\sigma(\Omega)$ is the interaction cross-section and we obtain Boltzmann's famous transport equation. Conservation laws of transport processes can be derived if all terms of Boltzmann's equation weighted by a certain function $\Theta(\mathbf{r}, \mathbf{v})$ are averaged with respect to the velocity. Note that the corresponding average of the collision term is zero; see Huang [6] for further details. E.g. energy conservation can be derived if Θ is related to the kinetic energy $(1/2)mv^2$. Therefore a classical multi-body system can be described in a statistical manner by a distribution function $f(\mathbf{r}, \mathbf{v}, t)$ as a solution of Boltzmann's equation or by its moments that can be interpreted in a dynamical manner as conservation laws. The semi-classical transport theory of semiconductors can be developed quantitatively if a two-fluid model is used and distribution functions as well as Boltzmann's equations for electrons and holes are formulated and ad-hoc quantum mechanical assumptions will be added. If average processes are used we obtain the well-known van Roosbroeck equations or drift-diffusion model for the electron density n and hole density p of semiconductors (see e.g. van Roosbroeck [7], Selberherr [8])

$$\begin{aligned} \nabla \cdot (\varepsilon \nabla \varphi) &= -e(p - n + N_D^+ - N_A^-), \\ \frac{\partial n}{\partial t} &= \nabla \cdot (-\mu_n n \nabla \varphi + D_n \nabla n), \\ \frac{\partial p}{\partial t} &= \nabla \cdot (\mu_p p \nabla \varphi + D_p \nabla p) \end{aligned} \tag{6}$$

with $D_{n,p}$ as diffusion coefficients, $\mu_{n,p}$ as mobilities, and the donator density N_D^+ as well as the acceptor density N_A^-.

Unfortunately the relationship $(1/2)mv^2$ is not valid in quantum mechanics and a modification of the semi-classical Boltzmann equation is needed. There are different options to do this. In each case the drift term is modified.

A first modification of the Boltzmann equation was presented by Wigner in 1932 [9] who introduced a non-local potential V. Wigner transformed Boltzmann's equation with respect to the velocity v into the k–space. Then the drift term is replaced by a memory term; in the 1-D case Wigner's variation of the Boltzmann equation can be formulated as

$$\frac{\partial f_w}{\partial t} = -\frac{\hbar k}{m}\frac{\partial f_w}{\partial x} - \frac{1}{\hbar}\int \frac{dk'}{2\pi} V(x, k - k') f_w(x, k, t) - \left(\frac{\partial f_w}{\partial t}\right)_{\text{coll}}, \tag{7}$$

where $f_w = f_w(x, k, t)$. Wigner's approach can be used for studying transport processes with quantum corrections. A first implementation for 1-D cases was presented by Biegel *et al.* [10] in the program SQUADS but a 2-D version of this program is not trivial (see Biegel [11]). An alternative concept for a Boltzmann equation with quantum corrections was presented by Nordheim [12] and Uhlenbeck [13].

A drift-diffusion model with quantum corrections was e.g. presented by Ancona [14]. Based on ideas of Madelung [15] and Bohm [16] originally introduced for an alternative interpretation of quantum mechanics Ancona added a quantum corrected potential ("Madelung-Bohm potential"). This model is denoted as "density-gradient model". Ancona's approach can be derived also from a quantum corrected variant so-called hydrodynamic approach that became popular in semiconductor device simulation during the last few years. Further details about quantum corrected hydrodynamical equations for semiconductor devices and its applications can be found in the literature, see e.g. [17], [18], [19], and others.

Another approach for the derivation of a quantum Boltzmann equation was given by Mahan [20] where the drift term was corrected, too. Mahan considered energy and impulse as independent variables and the modified distribution function is a solution of the following equation

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_r f + \mathbf{F} \cdot \left(\frac{\nabla_v}{m} + \mathbf{v} \frac{\partial}{\partial \omega} \right) f = I(f), \tag{8}$$

with $I(f)$ as interaction term. Mahan's approach based on a certain Green function of a non-equilibrium system where the Keldysh formalism is used. An alternative derivation from first principles was presented recently by Prüstel [21]. His starting point is the Liouville-von-Neumann equation for the density matrix ρ (see e.g. Mathis *et al.* [22]) $\dot{\rho} = i\mathcal{L}\rho$, where $\mathcal{L}$ is the so-called Liouville operator. Prüstel applied a decomposition of the density matrix ρ into a relevant and an irrelevant part by means of a Kawasaki-Guntron projector $\mathcal{P}$ $\rho_{irr}(t) = \rho - \mathcal{P}\rho(t)$, where $\mathcal{P}\rho(t)$ is the relevant part and he ends up with a first order approximation which leads to an equation that can be interpreted as the quantum Boltzmann equation including the desired term $\mathbf{F} \cdot \mathbf{v}\partial/\partial\omega$; see Röpke [23].

At this point we emphasize that all variants of Boltzmann equations need classical or quantum mechanical equations for the microscopic dynamics. These equations are reversible in its nature. Since any type of Boltzmann equations is irreversible an additional technique is needed to randomize the dynamical equations. Boltzmann had already used a corresponding argument to derive his interaction term. A generalized form of Boltzmann's argument is known as Markovian limit. In his derivation of Mahan's quantum Boltzmann equation Prüstel cancelled the non-diagonal terms of the relevant observable and showed that it is equivalent with a Markovian limit where a decoherence time $\tau_{decoh} \sim 1/(kT)$ is introduced. Therefore decoherence is established by considering a certain subspace in the space of observables and consecutive cancelling of non-diagonal elements of the relevant observable. Decoherence aspects are discussed in the paper of Mathis *et al.* [22] where it is established by reducing the density matrix under consideration of a factor of the tensor

product describing the state space. For fully quantum transport processes in semiconductors the many-body Schrödinger equation has to be used that will be discussed in the following chapter.

1.2. Quantum Mechanical Transport

The ability of a semiconductor crystal to carry a macroscopic quantity like electric current is determined by the band diagram or the electronic spectrum of the crystal. The crystal lattice of a semiconductor is consisted of a large number of ionized atoms providing the electrons a very complex energy profile. For temperatures above $0K$ the lattice atoms move around (vibrate) their zero position causing a time dependent perturbation of the lattice structure. In a more detailed modelling, we have to include the fact that electrons interact with each other and with the lattice. In the Schrödinger picture of quantum mechanics, the equation of motion for electrons in a crystal is

$$-i\hbar\frac{\partial}{\partial t}\hat{\Psi}(\{\mathbf{r}_i, s_i\}\{\mathbf{R}_j\}; t) = H(t)\hat{\Psi}(\{\mathbf{r}_i, s_i\}\{\mathbf{R}_j\}; t). \tag{9}$$

Thereby $\mathbf{r}_i$ are the electronic and $\mathbf{R}_j$ the core coordinates.[1] s_i denotes the spin coordinate. $\hat{\Psi}(\{\mathbf{r}_i, s_i\}\{\mathbf{R}_j\}; t)$ is the complete wave function of the many-particle system and H the Hamilton-operator for the particular system of interest. The indices i and j determine the number of electrons and ionized atoms respectively in the crystal.

1.2.1. Time independent Schrödinger equation

We assume that the cores are spatially fixed in the system and the configuration of lattice atoms is time invariant. This leads to two simplifications for Eq. (9): Firstly, the complete electron state $\hat{\Psi}$ is a function only of the electron and spin coordinates and time

$$\hat{\Psi} = \hat{\Psi}(\{\mathbf{r}_i, s_i\}; t)$$

and secondly, Hamilton operator H is time-independent. For a (perfect) semiconductor crystal, the Hamilton operator is given by

$$H = \sum_{i=1}^{N}\left[-\frac{\hbar^2}{2m}\nabla_i^2 + v_{\text{ext}}(\mathbf{r}_i)\right] + \sum_{i<j}^{N}\frac{q^2}{\|\mathbf{r}_i - \mathbf{r}_j\|}. \tag{10}$$

[1] $\{\mathbf{r}_i, s_i\}$ and $\{\mathbf{R}_j\}$ represent the set of all electronic and atomic coordinates, e.g. $\{f_i\} = (f_1, f_2, \ldots f_i \ldots f_N)$.

The first term on the right hand side denotes the kinetic energy of each electron, the second term, the periodic lattice potential, represents the interaction of the electrons with the atoms. The last term in Eq. (10) gives the interaction between the electrons in the crystal.

The electron state $\hat{\Psi}(\{\mathbf{r}_i, s_i\}; t)$ can be written as product of a pure time dependent and spatial part $\hat{\Psi}(\{\mathbf{r}_i, s_i\}; t) = \Phi(t)\Psi(\{\mathbf{r}_i, s_i\})$, and the time and spatial dependent parts can be separated. The time dependent part $\Phi(t)$ holds

$$i\hbar \frac{d}{dt}\Phi(t) = E\Phi(t), \quad \text{with} \quad \Phi(t) = e^{-i\frac{E}{\hbar}t}, \tag{11}$$

whereby the constant due to the integration is neglected. Substituting the solution for $\Phi(t)$ into (9) then gives for the spatial dependent part $\Psi(\{\mathbf{r}_i, s_i\})$

$$H\Psi(\{\mathbf{r}_i, s_i\}) = E\Psi(\{\mathbf{r}_i, s_i\}), \tag{12}$$

which is the stationary Schrödinger equation. Eq. (12) is a eigenvalue problem with appropriate boundary conditions. All solutions $\Psi(\{\mathbf{r}_i, s_i\})$ are functions of the Hilbert space $\mathscr{H}$.

1.2.2. Single electron approximation

If we neglect the Fermi characteristic of electrons and use the Hartree approximation [24], the many-body problem is reduced to a formal single-particle problem. The electron-electron interaction is restricted to the Coulomb interaction. With the general definition of electron density

$$n(\mathbf{r}) = \sum_{i=1}^{N} \sum_{s} \psi_i^*(\mathbf{r}, s)\psi_i(\mathbf{r}, s) \tag{13}$$

we can rewrite the Coulomb interaction term and obtain

$$v_{\text{coul}}(\mathbf{r}) = \sum_{i=1}^{N} \int d^3r' \psi_i^*(\mathbf{r}') \frac{q^2}{\|\mathbf{r} - \mathbf{r}'\|} \psi_i(\mathbf{r}') = \int d^3r' \frac{q^2}{\|\mathbf{r} - \mathbf{r}'\|} n(\mathbf{r}'). \tag{14}$$

In conclusion one gets the classical potential energy for the direct interaction. The Hartree approximation can be described as a mean field theory, whereby a single electron is moving in potential due to presence of the other electrons, but is not interacting with them. The potential acting on the electrons is the same for every electron. The potential $v_{\text{coul}}(\mathbf{r}) = qU(\mathbf{r})$ itself can be calculated from the Poisson equation

$$\nabla \cdot (\epsilon(\mathbf{r})\nabla U(\mathbf{r})) = -qn(\mathbf{r}), \tag{15}$$

with $qn(\mathbf{r})$ as the charge density. The resulting single electron equation of motion in the Hartree approximation is given by[2]

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + v_{\text{ext}}(\mathbf{r}) + v_{\text{coul}}(\mathbf{r})\right]\psi_m(\mathbf{r}) = \varepsilon_m\psi_m(\mathbf{r}). \tag{16}$$

The solution of the Poisson Eq. (15) and the equation of motion (16) has to be obtained self-consistently. Both equation are cross coupled due to the potential $v_{\text{coul}} = qU$ and the electron density n, see Eq. (13). Note that a single particle Schrödinger equation can only describe pure coherent transport of electrons throughout the device. For any loss of coherence due to inelastic scattering we need a generalized modelling concept.

1.2.3. Effective mass equation

Due to the periodicity of the crystal lattice potential v_{ext} we can separate the solution of the crystal Schrödinger equation into a periodic (Bloch functions) and a non-periodic part (envelope functions)

$$\psi(\mathbf{r}) = \hat{\psi}(\mathbf{r})u(\mathbf{r}), \tag{17}$$

and the equation of motion (16) is simplified to the effective mass equation

$$\left[-\frac{\hbar^2}{2m^*}\nabla^2 + v_{\text{coul}}(\mathbf{r})\right]\hat{\psi}_m(\mathbf{r}) = \varepsilon_m(\mathbf{k})\hat{\psi}_m(\mathbf{r}), \tag{18}$$

whereby the single electron orbital $\hat{\psi}_m(\mathbf{r})$ has to be distinguished from $\psi(\mathbf{r})_m$ in Eq. (16). Roughly speaking, we can say that the effective mass replaces the periodic potential of the crystal lattice but still contains structural properties of the crystal. The corresponding dispersion relation for Eq. (18) depends on the particle type (electrons or light or heavy holes etc.). For example, for electrons at the conduction band edge the dispersion relation is usually parabolically approximated

$$\varepsilon_m(\mathbf{k}) = E_{c0} + \frac{\hbar^2}{2m^*}(k_x^2 + k_y^2 + k_z^2), \tag{19}$$

whereby $\mathbf{k} = (k_x, k_y, k_z)$ is the wave vector and E_{c0} the conduction band edge. All needed quantities characterizing transport processes, like electron density n and current density $\mathbf{J}$, can be directly computed from the envelope function $\hat{\psi}(\mathbf{r})$ and it is not necessary to calculate the actual orbital $\psi(\mathbf{r})$, see e.g. [25].

[2]The spin coordinate s_i can be neglected in the single electron picture, when no magnetic effects are considered.

1.2.4. Spatial dependent effective mass

We assume a quasi-one dimensional solid, consisting of different materials (like a metal-oxide-semiconductor structure). The spatial coordinates are arranged so that the transport is happening in the x-direction and the y, z plane is the transverse plane. The material is independent from the y, z coordinates and changes only in the transport direction and the effective mass is a function of x, e.g. $m^* = m^*(x)$. The Hamilton operator from corresponding Schrödinger equation $H\hat{\psi} = E\hat{\psi}$ can be separated into a transverse part H_T and a longitudinal part H_L

$$H \equiv H_T + H_L \quad \text{and} \quad \hat{\psi}(\mathbf{r}) = \psi(x) \cdot \varphi(y, z). \tag{20}$$

For a solid with a very large cross section area (effectively infinite cross section) any confining potential in the transverse direction can be neglected.[3] The solution for the transverse direction follows directly in terms of plane waves

$$\varphi(\mathbf{r}_\perp) = \frac{1}{\sqrt{S}} e^{i\mathbf{k}_\perp \cdot \mathbf{r}_\perp}. \tag{21}$$

$\mathbf{k}_\perp$ and $\mathbf{r}_\perp$ are both vectors in the $y - z$ plane. S is the transverse cross sectional area.[4] The spatial dependence of $m^*(x)$ is accounted with the assumption

$$\frac{\hbar^2}{2} \frac{\partial}{\partial x} \frac{1}{m^*(x)} \frac{\partial}{\partial x} \tag{22}$$

which is in fact still a Hermitian operator and the eigenvalues remain real. The usage of spatial dependent mass operator (22) was controversial (see e.g. [26]) but accepted in nowadays [27] and commonly used in transport modeling [28, 29].

The Schrödinger equation for a spatial dependent mass $m^*(x)$ is given by

$$\left[-\frac{\hbar^2}{2m^*(x)} \frac{\partial^2}{\partial \mathbf{r}_\perp^2} - \frac{\hbar^2}{2} \frac{\partial}{\partial x} \frac{1}{m^*(x)} \frac{\partial}{\partial x} + v_{\text{Coul}}(x) \right] \hat{\psi}(\mathbf{r}) = E\hat{\psi}(\mathbf{r}). \tag{23}$$

The spatial dependent effective mass Eq. (23) can be reduced to a one dimensional equation in transport direction

$$\left[-\frac{\hbar^2}{2} \frac{\partial}{\partial x} \frac{1}{m^*(x)} \frac{\partial}{\partial x} + v_{\text{Coul}}(x) + E_c(x) \right] \psi(x) = \left(E - \frac{\hbar^2 k_\perp^2}{2m^*(x)} \right) \psi(x). \tag{24}$$

[3] Free electron problem in transverse direction.

[4] Note that S cancels out, when we calculate any physical quantity like current density etc.

E is the total energy and $\varepsilon_\perp$ the transverse eigenenergy

$$E = \varepsilon_x + \varepsilon_\perp, \quad \varepsilon_\perp = \frac{\hbar^2 k_\perp^2}{2m^*(x)}. \tag{25}$$

Due to the energy conservation law one can write equivalently for (25)

$$E = \varepsilon_x^L + \frac{\hbar^2 k_\perp^2}{2m_L^*} = \varepsilon_x + \frac{\hbar^2 k_\perp^2}{2m^*(x)}, \tag{26}$$

whereby ε_x^L and m_L^* are respectively the longitudinal eigenenergy and the (constant) effective mass at the point $x = 0$, which would be a lead[5] (or contact) when we consider a real device. Thus (24) changes to

$$\left[-\frac{\hbar^2}{2}\frac{\partial}{\partial x}\frac{1}{m^*(x)}\frac{\partial}{\partial x} + v_{\text{eff}}(\mathbf{k}_\perp, x)\right]\psi(x) = \varepsilon_x^L \psi(x) \tag{27}$$

with an effective potential dependent on direction x and the transverse modes $\mathbf{k}_\perp$

$$v_{\text{eff}}(\mathbf{k}_\perp, x) = v_{\text{Coul}}(x) + E_c(x) + \frac{\hbar^2 k_\perp^2}{2m_L^*}\left(1 - \frac{m_L^*}{m^*(x)}\right). \tag{28}$$

The corresponding Poisson equation is given by

$$\frac{d}{dx}\epsilon(x)\frac{d}{dx}U(x) + q[N_d^+(x) - N_A^-(x) - n(U, x)] = 0, \tag{29}$$

with the densities of the acceptor $N_A^-(x)$ and donor ions $N_d^+(x)$, respectively, the electronic charge q and a spatial dependent permittivity $\epsilon(x)$, caused by the device structure.

As a short summary, the two basic equations for coherent quantum transport are given by Eq. (27) and (29). The two equations have to be solved self-consistently, since they are cross-coupled due to

$$n(x) = \sum_\alpha \|\psi_\alpha(x)\|^2 f_0(\varepsilon_\alpha - \mu) \tag{30}$$

and $v_{\text{coul}}(x) = qU(x)$. f_0 is the Fermi function (see Eq. (47)) and μ the Fermi level.

1.3. Solving Transport Equations

A main difficulty in calculating the electron transports is to find an adequate method to solve the either semi-classical or the quantum mechanical transport

[5]Indicated by the superscript and index "L".

equation with the appropriate boundary conditions. As already emphasized, an electronic device has at least two contacts and is therefore an open system with open boundary conditions. The remaining part of this sections gives a short review about important approaches to solve transport equations. Due to the importance in the past, we start with semi-classical transport models.

1.3.1. Boltzmann equation and Wigner function approach

The Boltzmann Eq. (4) gives us the balance-equation for the number of particles $N(t)$ being inside the volume element $d^3r d^3v$ of phase-space at the time t, see Eq. (1). They will be scattered into and out of this volume element and will be accelerated by external fields, e.g. the electric field $\mathbf{E}$. In thermodynamical equilibrium we are able to get exact solutions from the Boltzmann equation. The Drift-Diffusion model (see Eq. (5)), the momentum-expansion of the Boltzmann equation, leads us to the continuity equation and an equation for the current-density $\mathbf{j} = \mu n \nabla E_F$ with the mobility μ, the electron density n and the Quasi-Fermi level E_F.

Taking into account quantum mechanical effect Wigner introduced a function analog to the classical probability density and containing the classical limit, the Wigner-function f_W [9].

Let ψ_n be the eigenfunctions and ε_n the eigenenergies for our system, described by the Schrödinger Eq. (18). In the thermodynamical equilibrium the density-matrix is $\rho\left(\mathbf{r}, \mathbf{r}'\right) = \sum_n e^{\frac{E_n}{KT}} \psi_n\left(\mathbf{r}\right) \psi_n\left(\mathbf{r}'\right)$ and the Wigner-function can be expressed as the Fourier-transformed of this density-matrix in center-of-mass coordinates.

$$f_W\left(\mathbf{x}, \mathbf{p}\right) = \left(\frac{1}{\pi\hbar}\right)^3 \int e^{\frac{2i}{\hbar}\mathbf{p}\cdot y} \rho\left(\mathbf{x}+\mathbf{y}, \mathbf{x}-\mathbf{y}\right) d^3y \tag{31}$$

Now, as Wigner showed, it is possible to gain information about a quantum system without solving Eq. (18), i.e. without determining the eigenstates or eigenvalues of our system.

Ancona was able to derive the Density-Gradient-method using the Wigner function approach, see e.g. [14, 30]. His result can be seen in a kind of generalized drift-diffusion equations with an additional correction-term, taking into account quantum-mechanical effects in the lowest degree. The Density-Gradient-method is most often used in simulations of semiconductor devices.

Finding the solution of the Boltzmann equation is a difficult problem since the distribution function has six arguments in the three dimensional case. The most widely technique for evaluating the Boltzmann equation is the Monte Carlo method [18]. Using this method the Boltzmann equation is not solved directly, but one rather simulates the motion of classical electrons subjected to a combination of free flight motion and instantaneous random scattering

events. The distribution function is then estimated by statistical averages over long times or many particles. The velocity and the position of each particle is integrated over the time between two collisions take place. Other random values determine the particular scattering mechanism and the velocity of the electron after the collision. After the collision takes place, the free-flight motion of the electron is again integrated until the next collision occurs. This procedure is performed for all electrons in the chosen ensemble to evaluate the time evolution of the device.

The Monte Carlo method permits to include other physical effects, such as detailed energy-band structure, electron-electron interaction and a more detailed description of scattering events.

1.3.2. 1-Particle-Schrödinger equation – nextnano^3

nextnano^3 is a versatile simulation software tool mainly developed at the "Physik Department and Walter Schottky Insitut of TU München", see e.g. [31]. Besides the 3D simulation of pure quantum mechanical devices like quantum dots it is also capable to calculate one or higher dimensionally and fully quantum mechanically the transport in classical devices. nextnano^3 contains a self-consistent Schrödinger Poisson solver, whereby self-consistency is achieved by introducing a spatial dependent quasi Fermi levels $E_F(\mathbf{r})$. The solutions of the Schrödinger Eq. (18) are assumed as a superposition of plane waves and the energies of these solutions are well defined by the dispersion relation $E(\mathbf{k})$. But, first of all the many-band-kp-Schrödinger equation is solved completely to get a good approximation for the band structure. This gives the density of charges (n for electrons and p for holes) by weighting the exact quantum mechanical states with the local Fermi levels. The local Fermi levels are obtained from the global current-conservation $\nabla \cdot \mathbf{j}_{n,p} = 0$, whereby $\mathbf{j}_n = \mu n \nabla E_{F,n}(\mathbf{r})$ for electrons and $\mathbf{j}_p = \mu p \nabla E_{F,p}(\mathbf{r})$ for holes.

For a summary the method is divided in two parts. In the first step the quasi-Fermi level is hold constant while the Schrödinger- and Poisson equation is solved self-consistently to get the new potential and the new quantized states. In the second step the potential and the states are fixed while determining the new Quasi-Fermi level with the current equation. This loop has to be repeated until self-consistency is reached.

1.3.3. Scattering-Matrix-Approach

The scattering matrix represents the solution of the Schrödinger Eq. (18) for a sample that is connected to semi-infinite leads, see [32] or [33]. In this method, carrier transport is viewed as the transmission and reflection of carrier fluxes

within a semiconductor. The simulation domain is subdivided into thin slices (1D) or meshes (2D), so that these regions are sufficiently small to assume constant doping and fields within. Transport across each region is described by a matrix equation which relates the incident carrier fluxes to the emerging fluxes, through the transmission coefficient of a scattering matrix. The potential term in the Schrödinger Eq. (18) is augmented with an additional potential $V(\mathbf{r})$ representing any impurities. The boundary conditions are chosen in a way so that the wave function vanishes outside of the sample and the leads. In this scheme the solution of our problem is a combination of plane waves moving towards and from the sample.The wave functions are normalized such that they carry unit flux. Inside a sample the solution of the Schrödinger equation is described by incoming ψ^i, outgoing waves ψ^o and evanescent waves, which are solutions with a complex wave vector. Far from the probe the evanescent mode will vanish. The scattering matrix can be divided into transmission and reflection matrices r, r', t and t'. For a wave approaching the sample through the left lead, the reflection matrix describes the reflected wave exiting through the left lead, and the transmission matrix t describes the transmitted wave in the right lead. Similarly, r' and t' describe reflection and transmission for waves coming from the right lead. Considering flux conservation demand the scattering matrix to be unitary.

For the case of serial scattering regions the description with transfer matrix can be used. This matrix relates the amplitudes in the left to the right of the sample. For the transfer matrices serial processes are expressed to be multiplicative. This multiplicative composition law points out the transfer matrix to be the ideal candidate for describing quantum transport through a disordered wire.

1.3.4. Pauli-Master-Equation

The method of using the Pauli-Master-equation for our transport problem has been pointed out by Fishetti *et al.* [34]. The principal in a shortcut: first of all solve the Schrödinger- and Poisson equation with eigenfunction approach and then use the eigenfunction as a basis for the quantum Liouville-equation and derive the Pauli-master equation. The Pauli-master equation throws out directly the occupation of states. The transition rates can now be calculated with the help of Fermi's-golden-rule. Let us have a quick look at this formalism.

Let $\psi_\eta(\mathbf{r})$ be a basis of the one-particle Hilbert space, describing our device and $\psi^n(\mathbf{r}) = \sum_\mu a^n_\mu(t)\psi_\mu(\mathbf{r})$ the state of our N-body-system at $t = 0$ we can use the density-matrix $\rho_{\mu\nu} = \sum_{n=1}^{N} a^n_\mu(t)a^{n*}_\nu(t)$ to write the Liouville equation

$$\frac{\partial\rho}{\partial t} = \frac{i}{\hbar}\mathcal{L}\rho + \left(\frac{\partial\rho}{\partial t}\right)_{\text{reservoir}} - \frac{\rho - \rho^{eq}}{\tau_s}. \tag{32}$$

In this context $\left(\frac{\partial \rho}{\partial t}\right)_{\text{reservoir}}$ takes the exchange with the reservoir into account, the third term realizes the influence of scattering inside our device with the scattering-time τ_s and the density in equilibrium ρ^{eq}. Using the basis-states of H_0 we reach the Pauli-Master equation and find the diagonal elements of the density-matrix ρ.

To use the sketched scheme for our transport problem we first of all have to solve the Schrödinger equation for an initial potential using mixed boundary conditions. We get bound-, left- and right-propagating states. The next is the calculation of transition-probabilities and the population of the states according to the Pauli-master-equation. The effect of the contact regions in our semiconductor device is phenomenologically expressed by the term $\left(\frac{\partial \rho}{\partial t}\right)_{\text{reservoir}}$. Each contact is mapped with a quasi-fermi-level, which has to be fitted to ensure the charge neutrality and current conservation in this region while performing the self-consistent loop. Hereafter the Poisson equation should be solved etc. Note that the Pauli-Master-Equation approach covers in general both coherent and incoherent transport.

1.3.5. NEGF formalism

A more sophisticated approach to quantum transport theory is supplied by the Green's function formulation of many-body theory. The non equilibrium Green's function (NEGF) theory was formulated by Kadanof and Baym in 1962 [35]. The application of non equilibrium Green's functions for the calculation of mesoscopic transport processes was mainly advanced by the group around S. Datta, R. Lake and M. Lundstrom [29, 36] and the NEGF is also the basis of quantum simulator `nemo`, see [37]. A very good introduction to the power of the NEGF formalism can be found in the two books [33] and [38] by S. Datta.

Roughly summarized, the non equilibrium Green's functions are defined by the expectation values of single-particle creation and annihilation operators. They describe the time evolution of the system. The Green's function is found by solving the Dyson equation, which is an integrated variant of the Schrödinger equation. The application of the NEGF to a MOS structure on the basis of the single electron Schrödinger equation will be shown in more detail in the next two sections. The presented example will exhibit stationary coherent transport, since this simplifications are appropriate for our aim, the analysis of tunneling currents in circuit simulations. However, the NEGF can be extended to include incoherent transport aspects as well as time dependent phenomena, see [35, 39, 40] and [33].

2. Self-consistent Transport Modeling

Modeling an electronic device as an open system means that we have to deal system of infinite extent. Thinking of a numerical calculation or simulation of such a system, means discretizing this infinite system and we would obtain matrices of infinite size, which would be intractable. The non-equilibrium Green's functions formalism (NEGF) offers a model for such open system, whereby the corresponding matrices are of finite extend (covering only the device region) and the coupling to the open environment is included in the finite discrete system. This section gives the most important conceptional steps of the NEGF and shows the calculation of coherent transport through multi-layered semiconductor structure. As we will see, the NEGF does not solve the Schrödinger equation directly but it calculates adequate quantities, including the needed electron density and the current density. The usage of a NEGF formalism to estimate tunneling currents might seem to much effort, in this case the Scattering matrix and the Transmission formalism would be sufficient. But our decision for the NEGF is explained by the versatility of this approach and the long term aim to extend considerations to incoherent and time dependent phenomena in charge transport.

Although all equations in the following have to be seen as discretized, e.g. differential equations change to matrix equations, we skip all considerations about discretization to the next section.

2.1. Green's Function for Coupled Device

For simplicity, we consider a system consisting of a device of finite dimension in x direction and with very large extend in transverse (y, z) direction and two semi-infinite electron reservoirs (contacts). The device is coupled to the two contacts, whereby the contacts are independent from each other. The Schrödinger equation for a isolated contact (i.e. contact 1) is given by

$$\big[\varepsilon_1 - H_1(x)\big]\psi_1(x) = 0. \tag{33}$$

We modify this equation to couple the isolated system with the environment and write

$$\big[(E + i\eta)\mathbf{1} - H_1(x)\big]\psi_1(x) = \mathscr{S}_1, \quad \eta \to 0, \tag{34}$$

whereby E is the independent energy variable and not the eigenvalue of the system. The term $i\eta\psi_1$ can be read as the extraction of the electrons from the contact and $\mathscr{S}_1$ as the re-injection of electrons from externals sources [38]. The extraction and re-injection of electrons keep the systems in equilibrium with its surroundings and maintains a constant electro chemical potential, see [38]

for further details. It is important to note that Eq. (34) is not a Schrödinger equation, but still gives the dynamics of a coupled system.

For the coupled contact-device-contact structure we obtain an equation describing the dynamics of the coupled system

$$\begin{bmatrix} (E+i\eta)\mathbf{1}-H_1 & -\tau_1^+ & 0 \\ -\tau_1 & E\mathbf{1}-H_d & -\tau_2 \\ 0 & -\tau_2^+ & (E+i\eta)\mathbf{1}-H_2 \end{bmatrix} \times \begin{bmatrix} \psi_1+\chi_1 \\ \psi_d \\ \psi_2+\chi_2 \end{bmatrix} = \begin{bmatrix} \mathscr{S}_1 \\ 0 \\ \mathscr{S}_2 \end{bmatrix}, \tag{35}$$

whereby $\mathbf{1}$ is the appropriate unity operator to maintain mathematical correctness. The wave functions for the contacts are divided into an incident part $\psi_{1,2}$ (also corresponding to the waveform of the isolated contacts 1,2) and a reflected waveform $\chi_{1,2}$. With Eq. (35) and (34) we can write

$$\big[(E+i\eta)\mathbf{1}-H_1\big]\chi_1 - \tau_1^+\psi_d = 0. \tag{36}$$

And the reflected waveform χ_1 in contact 1 can be estimated with

$$\chi_1 = g_1\tau_1^+\psi_d, \tag{37}$$

i.e., the reflected waveform is a response due an excitation ψ_d in the coupled device, whereby

$$g_1 = \big[(E+i\eta)\mathbf{1}-H_1\big]^{-1} \tag{38}$$

is the resolvent for the isolated contact. A corresponding expression can also be derived for contact 2

$$\chi_2 = g_2\tau_2^+\psi_d, \quad g_2 = \big[(E+i\eta)\mathbf{1}-H_2\big]^{-1}. \tag{39}$$

To described transport processes in the device, we need to estimate the Green's function G_d for the coupled device. From Eq. (35) follows

$$\big[E\mathbf{1}-H_d\big]\psi_d - \tau_1\chi_1 - \tau_2\chi_2 = \tau_1\psi_1 + \tau_2\psi_2. \tag{40}$$

With Eq. (37) $\chi_{1,2}$ can be substituted, which gives

$$\big[E\mathbf{1}-H_d-\Sigma_1-\Sigma_2\big]\psi_d = S. \tag{41}$$

Hence, the Green's function G_d for the coupled device is given by

$$G_d = [E\mathbf{1}-H_d-\Sigma_1-\Sigma_2]^{-1}. \tag{42}$$

and the wave function of the device is given by

$$\psi_d = G_d S. \tag{43}$$

The additional terms $\tau_1^+ g_1 \tau_1$ and $\tau_2^+ g_2 \tau_2$ incorporate the coupling of the finite device to the semi-infinite contacts. Both terms are called self-energies, defined by

$$\Sigma_1 \equiv \tau_1 g_1 \tau_1^+, \quad \Sigma_2 \equiv \tau_2 g_2 \tau_2^+. \tag{44}$$

For semi-infinite contacts, regular shaped and with well-defined transverse modes, the self-energies can be calculated analytically [33]. The source term S with a similarly meaning like $\mathscr{S}_{1,2}$ is defined by

$$S = S_1 + S_2, \quad S_{1,2} = \tau_{1,2} \psi_{1,2}. \tag{45}$$

2.2. Electron Density

In the semi-classical approach the electron density in equilibrium is given by

$$n = \int_{E_c}^{\infty} N(E) f(E) dE, \tag{46}$$

see e.g. [41] or [42]. $N(E)$ is the density of states, i.e., $N(E)dE$ gives the number of states in the interval $[E, E + dE]$. $f(E)$ denotes the statistical distribution function, which is in case of electrons the Fermi-Dirac distribution function,[6] since electrons are fermions. The Fermi function is defined as

$$f_0(E, \mu) = \frac{1}{1 + e^{\frac{E-\mu}{kT}}}, \tag{47}$$

whereby μ is the Fermi energy, which is usually obtained from the charge neutrality condition [42].

In the NEGF formalism the density of states can be obtained from the spectral function $A(E)$, which is defined by

$$A(E) \equiv i\big(G(E) - G^+(E)\big), \tag{48}$$

and can be seen as a more generalized concept of the density of states. The density of states is given by

$$N(E) = \sum_{\alpha} \delta(E - \varepsilon_\alpha), \tag{49}$$

and we can write for $A(E)$

$$A(x, x'; E) = \sum_{\alpha} \psi_\alpha(x) \delta(E - \varepsilon_\alpha) \psi_\alpha^*(x'). \tag{50}$$

[6] or usually shortened as the Fermi function.

$G(E)$ in Eq. (48) is the Green's function of the system.[7] The spectral function for the coupled device is given with Eq. (48) by

$$A_d = i\left(G_d - G_d^+\right) = i\left(\frac{1}{E - H_d - \Sigma} - \frac{1}{E - H_d - \Sigma^+}\right), \tag{51}$$

which is identical to Eq. (50), when ψ_α corresponds to the state of the coupled device $\psi_{d,\alpha}$. Similarly to the "generalized" density of states $A(E)$ the density matrix ρ as "generalized" electron density is introduced [33, 38]. To simplify the derivation of ρ, we reduce the problem and consider only the electron density in the device, which is caused by incident waves from contact 1. The density matrix of the device ρ_d is then given by

$$\rho_d(x, x') = \sum_\alpha \psi_{d,\alpha}(x) f_0(\varepsilon_\alpha - \mu_1)\psi^*_{d,\alpha}(x'), \tag{52}$$

whereby are the eigenstates of the isolated contact 1. Using the definition of $A(E)$ in Eq. (48) and the fact that $\psi_d = G_d S_1$ we obtain

$$\begin{aligned}
\rho_d &= \int f_0(E - \mu) \sum_\alpha \psi_{1,\alpha}\delta(E - \varepsilon_\alpha)\psi^* 1, \alpha dE \\
&= \int f_0(E - \mu) G_d \tau_1 \left[\sum_\alpha \psi_{1,\alpha}\delta(E - \varepsilon_\alpha)\psi^*_{1,\alpha}\right] \tau_1^+ G_d^+ dE \\
&= \frac{1}{2\pi}\int f_0(E - \mu) G_d \tau_1 a_1 \tau_1^+ G_d^+ dE.
\end{aligned} \tag{53}$$

with $a_1 = i[g_1 - g_1^+]$ the spectral function of the (isolated) contact 1. As discussed earlier, the coupling between contact and device is incorporated with the selfenergies $\Sigma_{1,2}$. Using this concept, broadening functions $\Gamma_{1,2}$ can be defined [38]

$$\Gamma_{1,2} \equiv i\left[\Sigma_{1,2} - \Sigma_{1,2}^+\right], \quad \text{and} \quad \Gamma_{1,2} = \tau_{1,2} a_{1,2} \tau_{1,2}^+. \tag{54}$$

The density matrix changes to

$$\rho_d(x, x') = \frac{1}{2\pi}\int_{E=-\infty}^{\infty} f(E, \mu_1) A_1 \, dE \tag{55}$$

with

$$A_1 \equiv G_d \Gamma_1 G_d^+, \tag{56}$$

[7] In case of our coupled device $G(E)$ would comply $G_d(E)$ from Eq. (42). $G^+(E)$ is the Hermitian conjugate of $G(E)$ and corresponds to the advanced Green's function of the system, while $G(E)$ is the retarded Green's function.

see [33] and [43]. The density matrix for the complete coupled system is simply the sum over all contacts [44]

$$\rho_d(x,x') = \int_{E=-\infty}^{\infty} \Big(f(E,\mu_1)A_1 + f(E,\mu_2)A_2 \Big)\, dE. \tag{57}$$

The electron density $n(x)$ is the diagonal of the density matrix

$$n(x) = \frac{1}{\Omega}\rho(x,x')|_{x=x'}, \tag{58}$$

whereby Ω is the volume of the unit cell.

2.3. Current Density

According to the continuity equation, the electrical current is given by

$$I = -q\frac{\partial n}{\partial t}. \tag{59}$$

The probability current therefore holds

$$I_p \equiv \frac{\partial}{\partial t}\big(\sum_{\alpha} |\psi_d(x)|^2\big). \tag{60}$$

The trace operation is identical with taking the summing over all α and we can write

$$\sum_{\alpha} |\psi_{d,\alpha}(x)|^2 = \sum_{\alpha} \psi^*_{d,\alpha}\psi_{d,\alpha} = \mathrm{Tr}\left[\psi_d^+\psi_d\right], \tag{61}$$

and obtain (see [38])

$$I_p = \frac{\partial}{\partial t}\big(\mathrm{Tr}\left[\psi_d^+\psi_d\right]\big). \tag{62}$$

The total (probability) current is zero, since we consider a non-equilibrium situation, caused by differing chemical potentials in the coupled reservoirs, but the situation is steady state. This means, all current going into the device, caused by contact 1 trying to bring the device in equilibrium with reservoir 1, is going out at contact 2, tyring to establish equilibrium with reservoir 2. Hence $I_1 = I_2 \equiv I$ and for the derivation of the current relation we only need to consider the current I_1 between contact 1 and device. The corresponding time dependent equation for the coupled system is given by

$$i\hbar\frac{\partial}{\partial t}\begin{bmatrix} \psi_1 + \chi_1 \\ \psi_d \end{bmatrix} = \begin{bmatrix} H_1 - i\eta & \tau_1^+ \\ \tau_1 & H_d \end{bmatrix}\begin{bmatrix} \psi_1 + \chi_1 \\ \psi_d \end{bmatrix}. \tag{63}$$

From Eq. (63) follows for the current I_1

$$I_1 = \frac{1}{i\hbar}\,\mathrm{Tr}\left[\psi_d^+\tau_1(\psi_1+\chi_1)+(\psi_1^+ + \chi_1^+)\tau_1^+\psi_d^+\right], \tag{64}$$

which can be divided into two parts. One corresponding to the incoming component, connected with the incident wave function ψ_1. The second part is the outgoing component, corresponding to the reflected wave function χ_1. Hence, we can write

$$I_1 = \underbrace{\frac{1}{i\hbar}\,\mathrm{Tr}\left[\psi_d^+\tau_1\psi_1+\psi_1^+\tau_1^+\psi_d^+\right]}_{\text{inflow}} - \underbrace{\frac{1}{i\hbar}\,\mathrm{Tr}\left[\chi_1^+\tau_1^+\psi_d+\psi_d^+\tau_1\chi_1^+\right]}_{\text{outflow}}. \tag{65}$$

With the substitution $\psi_d = G_d S$, whereby $S = S_1 + S_2$, the inflow component of the current can be formulated as

$$I_{1,\mathrm{in}} = \frac{1}{i\hbar}\,\mathrm{Tr}\left[S^+G_d^+S_1 - S_1^+G_dS\right] = \frac{1}{i\hbar}\,\mathrm{Tr}\left[S_1S_1^+G_d^+ - S_1S_1^+G_d\right], \tag{66}$$

whereby we used $S_1^+S_2 = S_2^+S_1 = 0$. The definition of the spectral function was given in Eq. (48), and the inflowing current in contact 1 reduces to

$$I_{1,\mathrm{in}} = \frac{1}{\hbar}\,\mathrm{Tr}\left[S_1S_1^+A_d\right]. \tag{67}$$

For the isolated contact 1 we can write according to Eq. (55)

$$\rho_1(x,x') = \psi_1\psi_1^+ = \int \frac{f_1(E)}{2\pi}a_1(E)dE. \tag{68}$$

Since $S_1S_1^+ = \tau_1\psi_1\psi_1^+\tau_1^+$ we obtain the expression

$$S_1S_1^+ = \int \frac{f_1(E)}{2\pi}\tau_1a_1(E)\tau_1^+dE = \int \frac{f_1(E)}{2\pi}\Gamma_1dE. \tag{69}$$

The inflow current is then given by

$$I_{1,\mathrm{in}} = \frac{1}{2\pi\hbar}\int f_1(E)\,\mathrm{Tr}\left[\Gamma_1A_d\right]. \tag{70}$$

The outflowing component can be derived similarly like the inflow. We substitute the reflected wave functions $\chi_1 = g_1\tau_1^+\psi_d$ and $\chi_1^+ = \psi_d^+\tau_1g_1^+$, and write

$$I_{1,\mathrm{out}} = \frac{1}{i\hbar}\,\mathrm{Tr}\left[\chi_1^+\tau_1\psi_d+\psi_d^+\tau_1^+\chi_1^+\right] = \frac{1}{i\hbar}\,\mathrm{Tr}\left[\psi_d\psi_d^+\Gamma_1\right]. \tag{71}$$

For $\psi_d\psi_d^+$ we can write

$$\psi_d\psi_d^+ = \rho_d = \int \frac{1}{2\pi}G^n dE, \tag{72}$$

whereby $G^n = f_1(E)A_1 + f_2(E)A_2$ is the electron correlation function, see Eq. (57). Therefore, the outflow component is written with

$$I_{1,\text{out}} = \frac{1}{2\pi\hbar}\int \text{Tr}\left[\Gamma_1 G^n\right] dE. \tag{73}$$

The total (electrical) current I is given with

$$I_1 = -2\frac{q}{h}\int (f_1 - f_2)\,\text{Tr}\left[\Gamma_1 G_d \Gamma_2 G_d^+\right] dE. \tag{74}$$

For further details see [38].

3. Numerical Transport Simulation

For an illustration of the numerical implementation of the NEGF formalism we consider a device structure consisting of two n-type silicon areas sandwiching an insulating oxide layer. The Si-oxide-Si device is connected on both sides to a contact. The complete device is pictured in Figure 1. In non-equilibrium, meaning the device is biased and the two Fermi levels $\mu_{1,2}$ of the left and right side reservoirs differ, a current flows through the structure and the insulating oxide layer. The considered situation is comparable to edge-direct-tunneling currents in modern MOSFET devices.

The following section starts with a discussion about the discretization of equation of motion and the calculation of all quantities of the NEGF formalism. Subsequent to the computation, we discuss the numerical solution of the Poisson equation, which we left out so far.

3.1. Method of Finite Differences

The calculation of physical quantities like currents and electron densities in devices is usually done with numerical simulations. This means, one has to

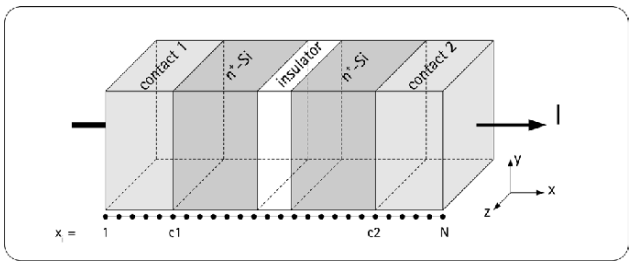


Figure 1. Si-oxide-Si device structure.

find a suitable discretization scheme for the underlying differential equations. In general, the equation of motion in the NEGF formalism and the coupled Poisson equation are discretized using the method of finite differences [29, 43, 45]. However, there also alternative discretization schemes to find in recent literature like the application of the finite element method in [46].

The 1-dimensional equation of motion for Si-oxide-Si structure with transport in x-direction (see Figure 1) is given by Eq. (27) on page 218. The finite difference form for (27) is written with

$$-t_{i-1,i}\psi_{i-1} + H_{d,i}\psi_i - t_{i,i+1}\psi_{i+1} = \varepsilon\psi_i, \tag{75}$$

with

$$H_{d,i} = \left[\frac{\hbar^2}{2s^2}\left(\frac{1}{m^-} + \frac{1}{m^+}\right) + E_{c,i} + v_{\text{coul},i} + \varepsilon_\perp\left(\frac{m_L^*}{m_i^*} - 1\right)\right]. \tag{76}$$

The notation f_i should be read as $f_i = f(x_i)$, whereby x_i is the i-th discrete spatial point. The lattice spacing s, e.g. the distance between two discrete points, is equidistant. Furthermore we have

$$m^- = \frac{m_{i-1} + m_i}{2}, \quad m^+ = \frac{m_i + m_{i+1}}{2} \quad \text{and} \quad t_{i,j} = \frac{\hbar^2}{(m_i + m_j)s^2}. \tag{77}$$

Hence the differential Eq. (27) changes to a matrix equation. As one can see in Eq. (75), an arbitrary lattice point i is coupled only to it's nearest neighbors with $t_{i,j}$, thus the finite difference approximation is a tight-binding model [29, 38]. Let us take look at the dispersion relation for the discretized device. In the contact region of the example (see Figure 1) the effective mass m_{x_i} with[8] $i = 1, 2 \ldots c1, c2 \ldots N$ is constant and equals m_L^*, obviously. The discretized equations of motion in this part of the device can be simplified to a equation with a constant effective mass, see (18) and the discrete dispersion relation is given with

$$\varepsilon = v_{\text{coul},i} + E_{c,i} + 2t(1 - \cos(ks)). \tag{78}$$

For very small $k \cdot s$, meaning the transition from the discrete to continuous case, the dispersion relation reduces to the parabolic band approximation. The open boundary conditions for the device are incorporated with the selfenergies, as stated in the previous section. For an 1-dimensional device, the derivation can be obtained following simply arguments as presented by Datta in [43]. From the matrix representation (75) we found that any lattice point couples only with its two direct neighbours (in the 1-dimensional case). This means that only the first point an the N-th point have to be "coupled" with the selfenergies to the

[8]Therefore, $i = c_1 + 1 \ldots c_2 - 1$ determines the Si-oxide-Si part of the device displayed in Figure 1.

reservoirs (semi-infinite contacts) on both sides. Hence, the selfenergies $\Sigma_{1,2}$ are matrices completely filled with zeros except one point:

$$\Sigma_1(1,1) = -\tilde{t}e^{jk_1 s}, \quad \Sigma_2(N,N) = \tilde{t}e^{jk_N s}. \tag{79}$$

The needed wave vectors k_1, k_N can be estimated from the dispersion relation in Eq. (78).

With the above stated equations we are able to calculate all necessary quantities for the NEGF formalism, the discrete Green's function G_d for the device

$$G_d(E,k_\perp) = \left[E\mathbf{1} - H_d - \Sigma_1 - \Sigma_2\right]^{-1}, \tag{80}$$

the spectral functions

$$A_{1,2}(E,k_\perp) = G_d(E,k_\perp)\Gamma_{1,2}(E)G_d^+(E,k_\perp), \tag{81}$$

and the broadening functions

$$\Gamma_{1,2}(E) = i\Big(\Sigma_{1,2}(E) - \Sigma_{1,2}^+(E)\Big). \tag{82}$$

The density matrix follows as an integral over the interested energy interval and the sum over all transversal $\mathbf{k}_\perp$-states

$$\begin{aligned}\rho = \frac{1}{2\pi}\sum_{k_\perp}\int_{E=-\infty}^{\infty} \Big[&f_0(E,\varepsilon_{k_\perp},\mu_1)A_1(k_\perp,E) \\ &+ f_0(E,\varepsilon_{k_\perp},\mu_2)A_2(k_\perp,E)\Big]dE\end{aligned} \tag{83}$$

with the Fermi functions given in (47) and the transversal eigenstates from Eq. (26). We can rewrite all equations to be explicit dependent on the transversal wave vector $\mathbf{k}_\perp$ instead of the transversal eigenstate $\varepsilon_\perp$. Using the periodic boundary conditions, the summation over all transverse wave vectors changes to an integral

$$\sum_{\mathbf{k}_\perp} \to \int d^2\mathbf{k}_\perp \frac{S}{4\pi^2} = \frac{S}{4\pi^2}\int 2\pi k_\perp dk_\perp, \tag{84}$$

whereby S denotes the size of the transversal area. But it cancels out, when we calculate a real physical quantity like the electron density n. As we know from Eq. (58), the electron density at the discrete lattice point are the diagonal elements of the density matrix weighted by the volume $\Omega = S \cdot s$ of the discretized cell, see Eq. (58). Introducing a slightly changed density matrix ρ'

$$\begin{aligned}\rho' = \frac{1}{4\pi}\iint \Big[&f_0(E,k_\perp,\mu_1)A_1(k_\perp,E) + f_0(E,k_\perp,\mu_2)A_2(k_\perp,E)\Big] \\ &\times dE k_\perp dk_\perp,\end{aligned} \tag{85}$$

the electron density can be calculated with

$$n(x_i) = \frac{1}{s}\rho'(x_i,x_i')|_{x=x_i} \tag{86}$$

without the knowledge of S.

3.2. Solution of The Poisson Equation

The Poisson equation of the example device in Figure 1 is discretized on the same spatial lattice as the equation of motion and we can write

$$\frac{1}{s^2}\left(\epsilon_i^- U_{i-1} - (\epsilon_i^- + \epsilon_i^+)U_i + \epsilon_i^+ U_{i+1}\right) + q\left[N_{D_i}^+ - N_{A_i}^- - n_i\right] = 0 \quad (87)$$

with

$$\epsilon_i^- = \frac{\epsilon_{i-1} + \epsilon_i}{2}; \quad \epsilon_i^+ = \frac{\epsilon_{i+1} + \epsilon_i}{2}. \quad (88)$$

For the solution of the Poisson equation we use a the standard Newton-Raphson algorithm (see e.g. [47]). Eq. (87) is rewritten as a matrix equation. The boundary conditions follow directly from the applied bias over the structure. When V is the external voltage, the values of the first and the N-th point of the lattice are given by

$$\frac{1}{s^2 q}\left(\epsilon_1^- 0 - (\epsilon_1^- + \epsilon_1^+)U_1 + \epsilon_1^+ U_2\right) + \left[N_{D_1}^+ - N_{A_1}^- - n_1\right] = 0$$

$$\frac{1}{s^2 q}\left(\epsilon_N^- U_{N-1} - (\epsilon_N^- + \epsilon_N^+)U_N + \epsilon_N^+ V\right) + \left[N_{D_N}^+ - N_{A_N}^- - n_N\right] = 0.$$

Hence, the values at the boundary are fixed due to the applied voltage V. The solution of the Poisson equation is formulated as a problem of finding the roots of a discrete function F. For the i-th spatial point F_i is given by

$$F_i = \frac{1}{qs^2}\left(\epsilon_i^- U_{i-1} - (\epsilon_i^- + \epsilon_i^+)U_i + \epsilon_i^+ U_{i+1}\right) + N_{D_i}^+ - N_{A_i}^- - n_i \quad (89)$$

and we have to solve

$$\sum_j \frac{\partial F_i^m}{\partial U_j^m}\delta U_j^{m+1} = \mathscr{J}_f(F_i^m)\delta U_j^{m+1} = -F_i^m, \quad j = 1, 2 \ldots N, \quad (90)$$

whereby m denotes the iteration index and the sum of the Jacobian $\mathscr{J}_f$ on the left hand side runs over all N points of the discrete lattice. The corrected value of U is determined by

$$U^{m+1} = U^m + \delta U_j^{m+1}, \quad (91)$$

with δU_j^{m+1} given by

$$\delta U_j^{m+1} = -\mathscr{J}_f^{-1}(F_m)F_m, \quad (92)$$

according Eq. (90). To calculate the derivation $\partial n/\partial U$ for the Jacobian $\mathscr{J}_f$ we use the approximation given in [29]

$$\frac{\partial n}{\partial U} \approx q\frac{\partial n}{\partial E_f}. \quad (93)$$

With Eq. (85) and (86) we obtain

$$\frac{\partial n}{\partial E_f} \approx \frac{2q}{s}\int\frac{dE}{2\pi}\int\frac{d^2\mathbf{k}}{4\pi^2}\left[-\frac{\partial f_0(E,\mu_1)}{\partial E}A_1(k_\perp,E) - \frac{\partial f_0(E,\mu_2)}{\partial E}A_2(k_\perp,E)\right], \tag{94}$$

using the fact that the quasi Fermi level in contacts correspond to the chemical potentials $\mu_{1,2}$ in the coupled reservoirs 1 and 2, see [29]. The derivation of the Fermi function with respect to the Fermi level is given by

$$\frac{\partial f_{1,2}}{\partial E_f} = \frac{1}{kT}f_{1,2}(1-f_{1,2}). \tag{95}$$

3.3. Numerical Solution

After reaching convergence in the self-consistent solution of the equation of motion and the Poisson equation, the current density is calculated with Eq. (74). The complete coupled and self-consistent solution of the equation for the dynamics (the Schrödinger equation) and for the electro-statics (the Poisson equation) follows the flow chart given in Figure 2. The flowchart shows the calculation for a certain external bias V. For a complete $J-V$ characteristic of a device, the depicted procedure has to be repeated for every bias value V_n. The calculation of a $J-V$ device characteristic usually starts at equilibrium

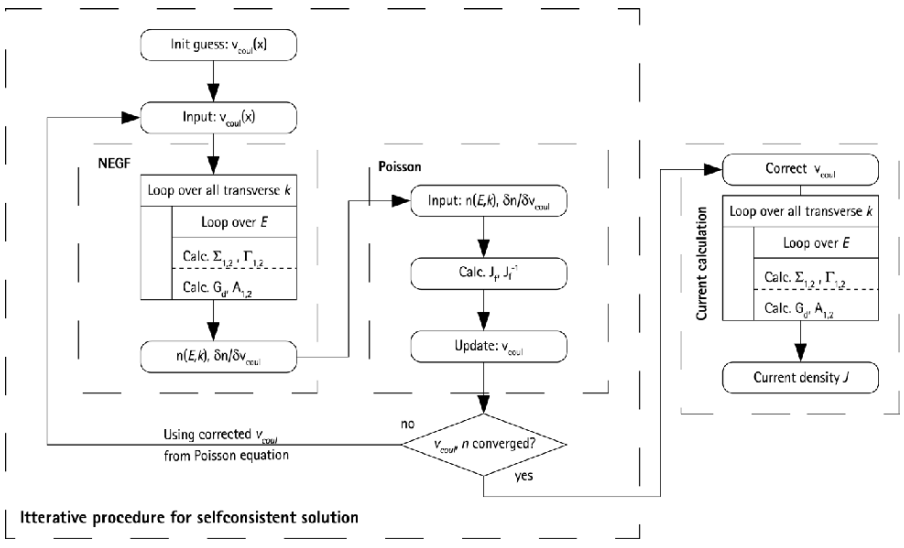


Figure 2. Flowchart of the iterative self-consistent solution for a certain external bias.

with $V = 0$, where a initial guess for U is easily done. For the calculation for different bias points, a reasonable initial guess for U is the converged U of preceding bias point, assuming that the difference between the bias points is not too big.

4. Circuit Simulation and Applications

We started our analysis of the impact of quantum effects to circuit functionality with the most obvious effect: the gate direct tunneling currents. The very first studies about the influence of tunneling currents can be found in the publications from Dutton, Choi *et al.* [48–50]. We use this publications as references for our overall methodology.

In modern MOSFETs two different direct gate tunneling mechanisms have to be accounted: (1) the direct tunneling between the inverted channel and the gate, and (2) the tunneling between overlapping source/drain extensions and the gate (edge-direct-tunneling). To include quantum parasitics in circuit simulations we represent them as additional Q-sources.[9] This means that in case of the direct tunneling currents in MOS-devices, such as FETs and capacitors, we are using additional voltage controlled current sources together with common device models. The current sources are implemented as look-up-table models, which is the easiest way to represent the current-voltage pairs from the numerical quantum transport simulations. Between two different pairs the values are obtained with linear interpolation, done by SPICE.

Our Q-sources are placed comparable to the tunneling leakage model implemented in BSIM4 (see [51]). The tunneling model in BSIM4 model differs significantly in estimating the magnitude of the tunneling currents. In comparison to our approach, the BSIM4 tunneling model needs a lot of non physical fitting parameters (see [51]), but which adjust the BSIM4 model to measurements of real physical devices. The withdraw of this very high-level or in other words non-physical description is that it covers the magnitude of macroscopic quantities (e.g. current density) for a particular device due to fitting. It is limited when the general behaviour is needed and when the dimension shrinks further and the quantum mechanical behaviour of the charge carriers causes more than additional currents.

4.1. SRAM Cell

In the recent literature, see e.g. [1, 52], the exponential increase of gate tunneling current at decreasing oxide thickness is a growing concern to ULSI

[9] Q $\hat{=}$ quantum parasitic.

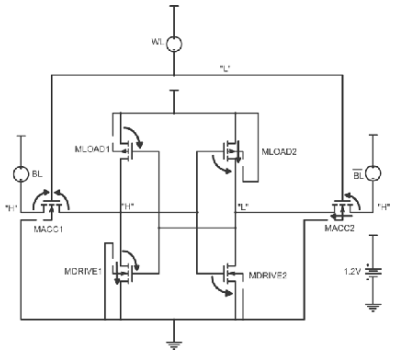


Figure 3. 6 transistor SRAM cell.

circuit performance and stability. But a direct influence of the quantum effects to the circuit functionality of static CMOS logic circuit itself is not expected to be a major problem [5]. In those cases only the strongly increasing off-state power dissipation is the main consequence [1]. As an example we simulated a 6-transistor SRAM cell [53] including our Q-sources. It showed that the ratio between the standard (or classical) leakage mechanisms and the direct tunneling leakage currents is reversing when the gate oxides gets thinner than 3 nm. The SRAM cell is depicted in Figure 3, whereby curved arrows show the tunneling currents and the straight arrows the conventional leakage currents. For a 2 nm oxide thickness flows total current of 9.2 pA caused by tunneling and the conventional leakage contributes only 2.9 pA to the power consumption. The simulation was made for the steady state situation depicted in Figure 3.

4.2. Domino-AND-2 Gate

In contrast to the "robustness" of static logic, dynamic logic and analog circuit functionality can be a critical case when the magnitude of tunneling current raises. It can be shown for a Domino-AND-2 gate that the circuit produces logical errors when the oxide thickness decreases beyond 2 nm [5]. The critical element in the Domino AND gate (see Figure 4) is the transistor M_2 at the input "A". In the precharge phase of the circuit (i.e. clock is "low", transistor M_1 is open) the capacitor C_1 is charged to V_{dd} level, so that the inverter at the output produces the correct "low" level (the corresponding input signal pattern is depicted in Figure 5). In the evaluation phase (clock signal is "high") the transistor M_1 is switched off and C_1 remains on V_{dd} as long the inputs "A" and "B" are zero. If we include the tunneling currents in the circuit simulation, we have an edge-direct tunneling in transistor M_2, which discharges C_1 as long the level on input "A" is "low". When the oxide thickness is under a critical value,

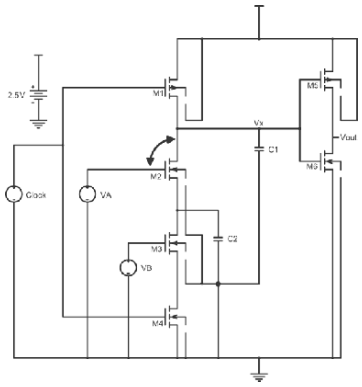


Figure 4. Domino AND 2 Gate.

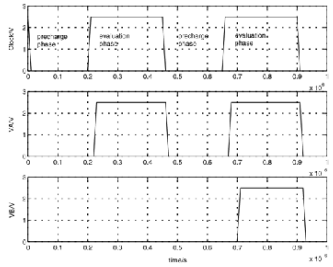


Figure 5. Input signals for Domino AND 2 Gate.

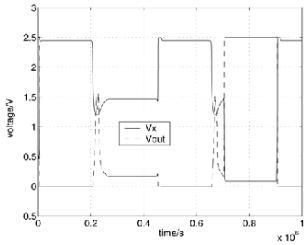


Figure 6. Output signals for Domino AND 2 Gate.

capacitor C_1 is discharged so quickly that V_x falls bellow $V_{dd}/2$ even before V_A on input "A" switches to a "high" level. In consequence, the inverter at the output produces a glitch with a magnitude higher than $V_{dd}/2$ (Figure 6), which has to be treated as a logical error – i.e. failure of the circuit functionality. A similar situation occurs when input "B" changes from "low" to "high" after a level change at input "A", see Figure 6.

4.3. Sample&Hold Circuit

The influence of gate direct tunneling currents in analog circuits can be shown with the example of a Sample & Hold circuit (suggested in [5]). The Spice schematic for a S&H circuit with a MOS capacitor as hold capacitance is shown in Figure 7. The switching transistors of the transmission gate and the MOS capacitor have a 2 nm oxide layer. In the transistor on-state the output waveform is directly following the input and the capacitance C_1 (see the Spice schematic in Figure 7) is charged to the current voltage level. With the charged MOS capacitor M_3 the output should remain on the last magnitude of the input when the transmission gate is switched off. But due to 2 nm insulating oxide, a direct tunneling current between the overlapping area of drain and gate and a direct tunneling in M_3 is discharging the output capacitance indicated be the arrows in Figure 7; and the output signal level is falling (see the dashed line in Figure 8) and does not remain as it should – circuit failure caused by Q-interference.

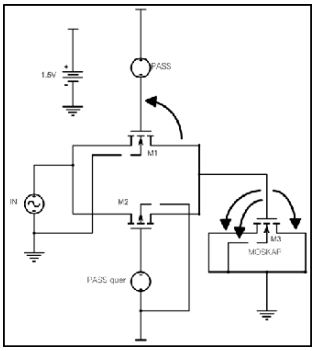


Figure 7. Sample & Hold circuit.

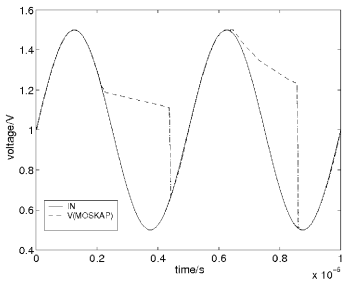


Figure 8. Input and output signals of a Sample & Hold circuit influenced by direct tunneling currents.

5. Conclusions

In this chapter, we tried to achieve two objectives: (1) to show that in future CMOS circuits, when length scale is in the domain around 20 nm and below, the charge carrier transport will be dominated by coherent transport and the drift-diffusion based device models will be too restricted and (2) to demonstrate that the functionality of classical circuits concepts can be substantially affected by parasitic quantum effects.

We showed a self-consistent NEGF-Schrödinger Poisson solver in detail, which allows the quantum mechanical calculation of coherent charge transport in semiconductors. Taking the results of the transport simulation we included parasitic quantum effects in high-level circuit simulators. Starting with the phenomena of direct tunneling currents in MOS circuits, we discussed different circuit examples, which all were affected in their functionality due to tunneling.

We believe among others [54] that the set of Schrödinger and Poisson equation will replace the various Drift-Diffusion models and hydrodynamic equations for device simulators in the <20 nm regime. Much work has to be done in order to maintain functionality of the important CMOS circuit concepts. More, former negligible quantum effects will disturb circuit functionality and the analysis of the influence of the tunneling currents can only be seen as a first step.

6. Acknowledgments

The Authors would like to thank Simon Fabel and Jan Bremer for their efforts to realise the numerical simulations.

References

[1] Roy, K.; Mukhopadhay, S.; Mahmoodi-Meimand, H. "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits", *Proc. IEEE*, **2003**, *91(2)*, 305–327.

[2] Yang, N.; Kirlen Henson, W.; Wortman, J.J. "A comparative study of gate direct tunneling and drain leakage currents in N-MOSFETs with sub-2-nm gate oxides", *IEEE Trans. Electron Dev.*, **2000**, *47(8)*, 1636–1644.

[3] Yu, Z.; Dutton, R.W.; Kiehl, R.A. "Circuit/device modeling at the quantum level", *IEEE Trans. Electron Dev.*, **2000**, *47(10)*, 1819–1825.

[4] Vasileska, D.; Knezevic, I.; Akis, R.; Ahmed, S.; Ferry, D.K. "The role of quantization effects on the operation of 50 nm MOSFETs, 250 nm FIBMOS devices and narrow-width SOI device structures", *J. Comput. Electron.*, **2002**, *1*, 453–465.

[5] Choi, C.-H.; Nam, K.-Y.; Yu, Z.; Dutton, R.W. "Impact of gate direct tunneling current on circuit performance: A simulation study", *IEEE Trans. Electron Dev.*, **2001**, *48(12)*, 2823–2829.

[6] Huang, K. *Statistical Mechanics*, John Wiley & Sons, **1963**.
[7] van Roosbroeck, W.V. "Theory of flow of electrons and holes in Germanium and other semiconductors", *Bell Syst. Techn. J.*, **1950**, *29*, 560–607.
[8] Selberherr, S. *Analysis and Simulation of Semiconductor Dev.*, Springer-Verlag, **1984**.
[9] Wigner, E. "On the quantum correction for thermodynamic equilibrium", *Phys. Rev.*, **1932**, *40*, 749–759.
[10] Plummer, J.; Biegel, B. "Comparison of selfconsistency iteration options for the Wigner function method of quantum device simulation", *Phys. Rev. B*, **1996**, *54*, 8070–8082.
[11] Biegel, B. "Simulation of ultra-small electronic devices: The classical-quantum transition region.", Technical Report, NASA, **1997**.
[12] Nordheim, L.W. *Proc. Roy. Soc London*, **1928**, *A 119*, 689.
[13] Uehling, E.A.; Uhlenbeck, G.E. "Transport phenomena in Einstein-Bose and Fermi-Dirac gases i.", *Phys. Rev.*, **1933**, *43*, 552–561.
[14] Ancona, M.G.; Iafrate, G.J. "Quantum correction to the equation of state of an electron gas in a semiconductor", *Phys. Rev. B*, **1989**, *39(13)*, 9536–9540.
[15] Madelung, E. "Quantentheorie in hydrodynamischer Form", *Zeitschr. Phys.*, **1927**, *40*, 322.
[16] Bohm, D. "A suggested interpretation of the quantum theory in terms of "hidden" variables i, ii.", *Phys. Rev.*, **1952**, *85*, 166–193.
[17] Jüngel, A. "Nonlinear problems in quantum semiconductor modeling", *Nonlinear Analysis*, **2001**, *47*, 5873–5884.
[18] Jungemann, C.; Meinerzhagen, B. *Hierarchical Device Simulation: The Monte-Carlo Perspective*, Springer-Verlag, **2003**.
[19] Ringhofer, C. "Computational methods for semiclassical and quantum transport in semiconductor devices", *Acta Numerica*, **1997**, *6*, 485–521.
[20] Mahan, G.D. "Quantum transport equation for electric and magnetic fields", *Phys. Rep.*, **1987**, *145*, 251.
[21] Prüstel, T. *Effektive Feldtheorien in Reeller Zeit. Master's thesis*, University of Hamburg, **2000**.
[22] Mathis, W.; Pahlke, K.; Zou, X.-B. "Decoherence in quantum systems and the network paradigm", *Intern. Journ. Circuit Theor. Appl.*, **2003**, *31*, 11–21.
[23] Röpke, G. *Statistische Mechanik für das Nichtgleichgewicht*, Physik-Verlag, **1987**.
[24] Bohm, A. *Quantum Mechanics – Foundations and Applications*, 3rd edition, Springer Verlag, **2001**.
[25] Datta, S. *Quantum Phenomena*, volume 8 of *Modular Series on Solid State Dev.*, 1st edition, Addison-Wesley, **1989**.
[26] Oldwig von Roos, "Position-dependent effective masses in semiconductor theory", *Phys. Rev. B*, **1983**, *27(12)*, 7547–7552.
[27] Levy-Leblond, J.-M. "Position-dependent effective mass and galilean invariance", *Phys. Rev. A*, **1995**, *52(3)*, 1845–1849.
[28] Einspruch, N.G.; Frensley, W.R. Eds., *Heterostrcutures and Quantum Dev.*, volume 24 of *VLSI Electronics: Microstructure Science*, Academic Press, **1994**.
[29] Lake, R.; Klimeck, G.; Chris Bowen, R.; Jovanovic, D. "Single and multiband modeling of quantum electron transport through layered semiconductor devices", *J. Appl. Phys.*, **1997**, *81(12)*, 7845–7869.
[30] Ancona, M.G. "Macroscopic description of quantum-mechanical tunneling", *Phys. Rev. B*, **1990**, *42(2)*, 1222–1233.

[31] Majewski, J.A.; Birner, S.; Trellakis, A.; Sabathil, M.; Vogl, P. "Advances in the theory of electronic structure of semiconductors", *Physica Status Solidi (c) 1(8), 2003 (2004)*, **2004**, *8*, 2003–2027.
[32] Ferry, D.K.; Goodnick, S.M. *Transport in Nanostructures*, Cambridge University Press, **1997**.
[33] Datta, S. *Electronic Transport in Mesoscopic Systems*, 5 edition, Cambridge Studies in Semiconductor Physics and Microelectronic Engineering. Cambridge University Press, **2003**.
[34] Fischetti, M.V. "Master-equation approach to the study of electronic transport in small semiconductor devices", *Phys. Rev. B*, **1999**, *59*, 4901–4917.
[35] Kadanoff, L.P.; Baym, G. *Quantum Statistical Mechanis*, W.A. Benjamin Inc., **1962**.
[36] Datta, S. "A simple kinetic equation for steady-state quantum transport", *J. Phys.: Condensed Matter*, **1990**, *2*, 8023–8052.
[37] Blanks, D.; Klimeck, G.; Lake, R.; Chris Bowen, R.; Frensley, W.R.; Leng, M.; Fernando, C.L. "Nanoelectronic modeling (nemo): A new quantum device simulator", Technical report, NASA Technical J., **1997**.
[38] Datta, S. *Quantum Transport – Atom to Transistor*, Cambridge University Press, **2005**.
[39] Haug, H.; Jauho, A.-P. *Quantum Kinetics in Transport and Optics of Semiconductors*, Springer Series in Solid-State Sciences. Springer Verlag, **1998**.
[40] Bayfield, J.E. *Quantum Evolution*, Wiley & Sons, Inc., **1999**.
[41] Pierret, R.F. *Advanced Semiconductor Fundamentals*, volume 6 of *Modular Series on Solid State Dev.*, 2nd edition, Prentice Hall, **2003**.
[42] Sze, S.M. *Physics of Semiconductor Dev.*, 2nd edition, Wiley-Interscience, **1981**.
[43] Datta, S. "Nanoscale device modeling: the Green's function method", *Superlattices Microstruct.*, **2000**, *28(4)*, 253–278.
[44] Paulsson, M. *Non Equilibrium Green's Functions for Dummies: Introduction to the One Particle NEGF Equations*, **2002**.
[45] Svizhenko, A.; Anantram, M.P.; Govindan, T.R.; Biegel, B. "Two-dimensional quantum mechanical modeling of nanotransistors", *J. Appl. Phys.*, **2002**, *91(4)* 2343–2354.
[46] Polizzi, E.; Datta, S. "Multidimensinal nanoscale device modeling: the finite element method applied to the non- equilibrium green's function formalism", in *IEEE NANO*, **2003**, 40–43,
[47] Hanke-Bourgeois, M. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Teubner, **2002**.
[48] Choi, C.H; Goo, J.-S.; Yu, Z.; Dutton, R.W.; Bayoumi, A.; Cao Paul Vande Voorde, M.; Vook, D. "C-V and gate tunneling current characterization of ultra-thin gate oxide MOS ($t_0x = 1.3 - 1.8$ nm)", in *Sympos. VLSI Technol. Digest of Technical Papers*, **1999**, xx–xx.
[49] Choi, C.-H.; Yu, Z.; Dutton, R.W. "Two-dimensional polysilicon quantum-mechanical effects in double-gate SOI", in *IEDM Technical Digest. Int.*, **2002**, 723–726.
[50] Choi C.-H.; Dutton, R.W. "Implications of gate tunneling and quantum effects on compact modeling in the gate-channel stack", in *The 2003 Nanotech Conf. Proc.*, **2003**, xx–xx.
[51] Liu, W. *MOSFET Models for SPICE Simulation including BSIM3v3 and BSIM4*, John Wiley & Sons, Inc., **2001**.
[52] Kirklen Henson, W.; Yang, N.; Kubicek, S.; Vogel, E.M.; Wortman, J.J.; De Meyer, K.; Naem, A. "Analysis of leakage currents and impact on off-state power consumption

for CMOS technology in the 100-nm regime", *IEEE Trans. Electron Dev.*, **2000**, *47(7)*, 1393–1400.

[53] Nii, K.; Tsukamoto, Y.; Yoshizawa, T.; Imaoka, S.; Yamagami, Y.; Suzuki, T.; Shibayama, A.; Makino, H.; Iwade, S. "A 90-nm low-power 32-kb embedded sram with gate leakage suppression circuit for mobile applications", *IEEE J. Solid-State Circuits*, **2004**, *39(4)*, 684–693.

[54] Jungemann, C.; Subba, N.; Goo, J.-S.; Riccobene, C.; Xiang, Q.; Meinerzhagen, B. "Investigation of strained Si/SiGe devices by MC simulation", *Solid-State Electron.*, **2004**, *48*, 1417–1422.

Chapter 9

COMPACT MODELING OF THE MOSFET IN VHDL-AMS

Christophe Lallement[1,a], François Pêcheux[2], Alain Vachoux[3] and Fabien Prégaldiny[1,b]

[1] *InESS (UMR 7163)/ ENSPS, Parc d'innovation, BP 10413, F-67412 Illkirch Cedex, France*

E-mail: [a] christophe.lallement@ensps.u-strasbg.fr,
[b] fabien.pregaldiny@iness.c-strasbourg.fr;

[2] *LIP6 Integrated Systems Architecture Department, Université Pierre et Marie Curie, 12, rue Cuvier, 75252 Paris Cedex 05, France*

E-mail: francois.pecheux@lip6.fr;

[3] *LSM Microelectronic Systems Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

E-mail: alain.vachoux@epfl.ch.

Abstract: In this chapter, we present the capabilities of the VHDL-AMS hardware description language for developing compact models. After a brief description of the VHDL-AMS language, we present two meaningful case studies on design oriented models of MOSFET.

The first study focuses on the EKV v2.6 MOSFET model and takes into account the thermo-electrical interaction and the extrinsic aspects. The EKV v2.6 model uses linearization with respect to surface potential, resulting in physically well-based expressions for the whole model.

The second study is a simplified version of the MM11 Philips model that takes into account the quantum mechanical effects. MM11 is a compact MOSFET model based on the formulation of the surface potential.

Key words: Hardware description language; VHDL-AMS; compact modeling; MOSFET model; thermo-electrical interactions; quantum effects; EKV; MM11.

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, 243–269.

1. Introduction

For the past three decades, hardware description languages (HDLs) have been widely used to model and simulate systems belonging to various engineering fields, from digital and analog electronics to mechanics and chemistry. For a long time, all these fields have been completely separated, each scientific community having its own design methodologies, tools and idiosyncrasies. For example, in the electrical/electronic domain, the SPICE simulator and all its derivatives allow the description of the netlist of a circuit using electrical primitives such as resistors, capacitors, sources and transistors. In an attempt to support the modeling and simulation of non-electrical systems as well, several modeling methods using energy equivalences between the electrical domain and other domains such as mechanical, thermal or fluidic domains have been proposed. With the advent of nano-technologies, the design of innovative integrated devices, like Micro-Opto-Electro-Mechanical Systems (MOEMS), has shifted from vertical only to both vertical and horizontal integration. Using the benefit of all the experience acquired in incremental design, MOEMS design now involves strong "horizontal" interaction of different application-field parts on the very same chip (e.g., mechanical, electrical, thermal, fluidic parts), with partial close coupling between these fields. Neglecting the interaction effects or the cross coupling between parts may have disastrous consequences on the final design in terms of a loss in performance or an increase in design time.

One way of addressing this issue is to use a consistent modeling and simulation framework that allows for the description of systems from different disciplines and for the description of interactions between these systems. This is where the VHDL-AMS HDL comes in action.

2. VHDL-AMS: A Mixed-Signal HDL

VHDL-AMS [1–4] is the result of an IEEE effort to extend the VHDL language to support the modeling and the simulation of analog and mixed-signal systems. The effort culminated in 1999 with the release of the IEEE standard 1076.1-1999 [1].

VHDL-AMS supports the description of continuous-time behavior. For compact modeling, the most interesting feature of the language is that it provides a notation for describing Differential Algebraic Equations (DAE) in a fairly general way. The "==" operator and the way unknown variables are declared allow the designer to write equations in either implicit or explicit format.

VHDL-AMS supports the description of networks as conservative-law networks (Kirchhoff networks) and signal-flow networks (inputs with infinite impedance, outputs with zero impedance). As such, it supports the description

and the simulation of multi-discipline systems at these two levels of abstraction. As a companion standard, the IEEE 1076.1.1-2004 standard includes packages that define types, subtypes, natures and constants for modeling in multiple energy domains [2].[1]

The VHDL-AMS language has a canonical, tool independent, mixed-signal simulation cycle that defines how to simulate a mixed-signal description. It has in addition a formal definition of how to initialize a mixed-signal model. It supports continuous-time analyses such as time-domain, DC, small-signal AC and noise analyses.

VHDL-AMS does not provide any support of the SPICE netlist format, neither directly in the language nor in some standard library. It however provides all the necessary language elements to build libraries of SPICE models. This is certainly helpful when developing compact models.

Any VHDL-AMS design unit may be compiled separately and stored in a library. In addition, VHDL-AMS allows for a clear separation between the interface of a model and its internal description and provides a mechanism to select the submodels to use in a hierarchical description through the mechanism of configuration. Both capabilities hence allow for much flexibility when it comes to model large complex hierarchical systems.

Table 1 presents a synthetic view of the capabilities offered by VHDL-AMS.

Table 1. Key Features of VHDL-AMS [7].

Features class	Feature	VHDL-AMS
Language aspects	Definition	IEEE Std 1076.1-1999 Strict extension to IEEE Std 1076 (VHDL)
	Inheritance	Ada-like. Case insensitive
	Modularity	Separation of external/interface views (entities) and internal views (architectures), packages, configurations
	Genericity	Parameters, generate statements(1)
	Library management	Yes (pre-compiled design units)
	Analog subset	No(2)
Expression of structure	Ports	Event-driven and continuous Conservative and non conservative (signal-flow) Continuous ports are modeless
Expression of behavior	Composition	Hierarchical instantiation of components

(*Contd.*)

[1]However, as the new standard is not yet fully supported in the current tools, we shall use proprietary versions of the packages. The proprietary packages do not actually differ much from the standard ones.

Table 1. Continued.

Features class	Feature	VHDL-AMS
	Conservative semantics	Natures define energy domains, subtypes define nature attributes; no predefined natures. Terminal and branch quantities
	Objects	Terminals, quantities, signals, variables, constants
	Statements	Concurrent, sequential, continuous (simultaneous and procedural) Continuous statements can be freely mixed with concurrent statements
	Expression of DAEs[3]	Explicit and implicit form of equations[4] supported Simultaneous[5] and procedural[6] formulations Derivative attribute 'dot only possible on quantities. Attribute can be chained for higher order derivatives[7] Mathematical functions defined in separate standard IEEE 1076.2 Piecewise defined behavior supported[8]
	Discontinuity handling	Discontinuities must be explicitly announced in the model User-defined re-initialization after discontinuity supported
Conservative semantics	Energy domains	Natures define energy domains and subtypes define nature attributes. No predefined natures[9] Branch quantities[10]
	Formulation	Equation-oriented formulation with simultaneous statements[11] No specific circuit graph representation enforced
Signal-flow semantics	Model interface	Directional interface (free) quantities[12]
	Functional blocks	Laplace and z transforms
Mixed-signal aspects	Interfaces	A/D and D/A interface language attributes ('ramp, 'slew, 'above) No direct port association[13]
	Behavioral interactions	Access of discrete signals in continuous context Access of continuous quantities in discrete context

(*Contd.*)

Table 1. Continued.

Features class	Feature	VHDL-AMS
Simulation controls	Solvability	Solvability check done at design unit level(14)
	Timestep	Timestep size may be bounded
	Tolerances	Generic string annotation not formally linked to simulator(15)

(1) Generate statements offer macro-like capabilities in the text of the model.
(2) It is possible to develop packages to support SPICE level modeling. No standard packages exist yet.
(3) DAE = Differential Algebraic Equation.
(4) An equation in the explicit form looks roughly like an assignment, e.g. $x = f(y, z)$, while an equation in the implicit form typically requires iterations to compute the unknowns, e.g. $x = f(x, y, z)$.
(5) Simultaneous statements are basically equations that may be given in any order in the model.
(6) Procedural statements have to be given in a particular order. The VHDL-AMS tool used did not support simultaneous procedural statements yet, so we used functions instead.
(7) To maintain good numerical accuracy it is recommended to hold higher order derivatives in local quantities and to only use first order derivatives.
(8) Continuous behavior can be defined by regions of operation.
(9) A draft VHDL-AMS standard package for multiple energy domain support is currently under IEEE ballot.
(10) The direction of the flow in the branch and of the potential difference is defined in the branch quantity declaration.
(11) Quantities are the unknowns. As far as the language is defined, the order in which the simultaneous statements is not important. We anyway faced some non-convergence issues with "misplaced" simultaneous statements (tool issue).
(12) Direction is used for solvability checks.
(13) It is not allowed to associate formal and actual ports of different natures or types. Explicit interface code has to be added in the model when pre-defined attributes are not enough. It is also expected that tools may help in inserting proper interface code when working at schematic level.
(14) This basically checks that the number of unknowns matches the number of equations. Although the rules that define what is considered as an unknown and what is considered as an equation are clearly defined in the language reference manual, it may become pretty hard to figure out what is missing when a complex model such as the full EKV MOS model (with more than 100 quantities) does not comply with the solvability condition. In addition, the current implementation of the Mentor tool imposes to have the same number of simple simultaneous statements in each branch of a conditional or selective simultaneous statement.
(15) Current VHDL-AMS simulators are using their own tolerances that may be set in the tool's environment.

For this chapter, the EDA tools used for implementing and simulating the models are Advance MS from Mentor Graphics and Simplorer from Ansoft.

To be complete, it should be noticed that VHDL-AMS has a direct competitor: Verilog-AMS [5–7]. The Verilog-AMS language also supports the modeling and the simulation of analog and mixed-signal systems but has not been submitted yet to IEEE for standardization [7].

3. Compact Modeling of the MOSFET

Compact models for circuit simulation have been at the heart of CAD tools for circuit design over the past decades, and are playing an ever increasingly important role in the nanometer system-on-chip (SoC) era. The requirements for a competitive compact MOSFET model rely on a complex trade-off between accuracy, complexity and applicability for any advanced technology. To achieve this task, in particular for devices entering the sub-100-nm regime, it is essential to accurately model the physical effects that govern the MOSFET behavior. This is the reason why a new generation of MOSFET models (the 4th one) is being developed (see Figure 1). Conventional models of the 3rd generation like BSIM3 / BSIM4 [8] and MM9 [9] are based on the formulation of the threshold

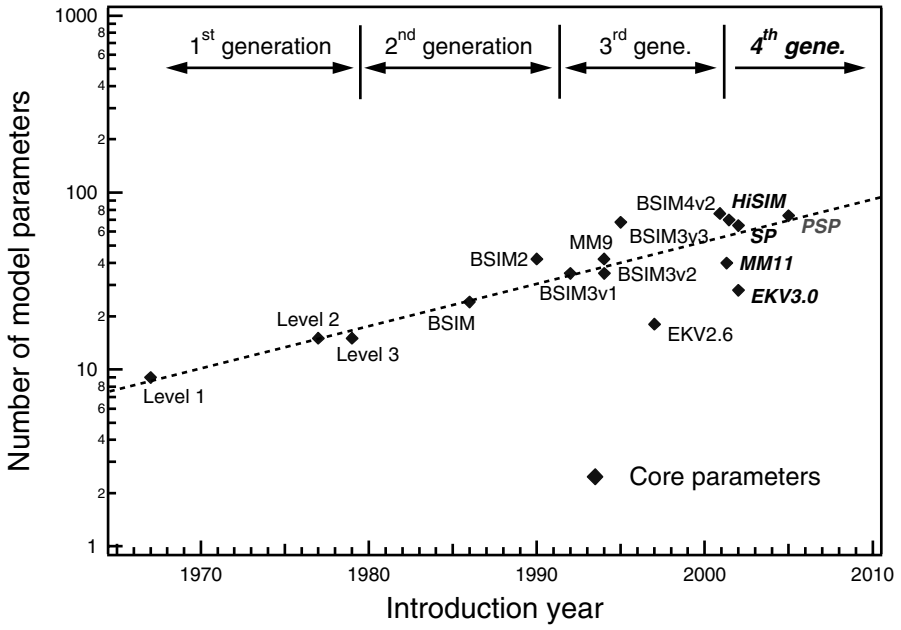


Figure 1. Number of core parameters (i.e., without geometrical or scaling parameters) for the major compact MOSFET models versus the introduction year of models.

voltage. The case of EKV v2.6 [10] is quite different since the model is based on the formulation of the inversion charge density. In fact, EKV v2.6 should already be considered as a 4th generation MOSFET model. For convenience, we use the term "inversion charge model" to refer to such a model. From the historical point of view, the major characteristics of the 3rd generation were [11]:

- the "original intent" to simplicity (in contrast to the 2nd generation models like BSIM and BSIM2),
- a small number of physically-based parameters,
- an improved mathematical conditioning,
- a single model equation for all regions of device operation,
- the use of smoothing functions.

Unfortunately, the most used model in the design community (BSIM3/4) has forgotten the original intent of both simplicity and small number of parameters, as depicted in Figure 1. All 3rd generation models (except EKV v2.6 which *is not* a threshold-voltage-based model) describe different operating regions with different equations. As a result, they are usually called "piece-wise" or "regional" models [11]. They often use unphysical parameters to smooth characteristics between the different operation modes. This artificial modeling may lead to unphysical behavior of the drain current and transconductance in the transition region between weak and strong inversion, the so-called moderate inversion region [12]. This region is however of crucial importance, not only for low-voltage and low-current analog applications, but also for digital circuits, owing to the reduction of supply voltage in modern CMOS technologies. Moreover, for most analog applications the device is typically biased in this region, i.e. just above threshold. Another drawback of regional models is that the drain current exhibits a discontinuity at the transition between linear and saturation regions due to the use of the drift approximation. Consequently, additional parameters are needed to get continuous characteristics through different operation modes, and the total number of parameters dramatically increases (see Figure 1).

In contrast to regional models, the compact models of the 4th generation like EKV 3.0 [13], HiSIM [14], MM11 [15], SP [16] and now PSP [17] are inherently single-piece and give an accurate and continuous description of characteristics in all regions of operation. They are generally charge sheet models based on the formulation of the surface potential, except EKV 3.0 which is a charge sheet model based on the formulation of the inversion charge density. Using the drift-diffusion approximation these models are more able to support future technology requirements.

In conclusion, and for the sake of completeness, it should be noted that a new compact MOSFET model called PSP is now available [17]. It has been developed by merging the best features of two surface-potential-based models: SP

(developed at The Pennsylvania State University) and MM11 (developed by Philips Research). The PSP model is a symmetrical model, and gives an accurate physical description of the transition from weak to strong inversion and includes an accurate description of all physical effects important for modern and future CMOS technologies. It is suitable for digital, analog and RF circuit design.

4. The EKV MOSFET Model v2.6

The EPFL EKV MOSFET model v2.6 is a scalable and compact simulation model built on fundamental physical properties of the MOS structure. This model is dedicated to the design and simulation of low-voltage, low-current analog, and mixed analog-digital circuits using submicron CMOS technologies.

4.1. Basic Version

The basic version of the EKV v2.6 MOSFET model [10] is a charge-based compact model. It consistently describes effects on charges, transcapacitances, drain current and transconductances in all regions of operation of the MOSFET (weak, moderate, strong inversion) as well as conduction to saturation. The effects modeled in this model include all the essential effects present in submicron technologies. For quasi-static dynamic operation, both a charge-based model for the node charges and transcapacitances, and a simpler capacitances model are available.

4.2. Features Specific to Submicron CMOS Technologies

The LDD regions in the sub- and deep- submicron CMOS technologies introduce additional parasitic resistances between the source/drain electrode and the channel, as well as parasitic capacitances.

A problem with all these parasitic elements is their non-linear and bias dependent behavior. An efficient MOSFET model dedicated to deep-submicron design must imperatively take into account these elements. These features specific to submicron technologies are not included in the basic version of the EKV MOSFET model v2.6. They have been added to the basic code of the EKV MOSFET model v2.6, in a modified version of this model, implemented in VHDL-AMS.[2]

[2]The full VHDL-AMS code of the model is available on a website [20].

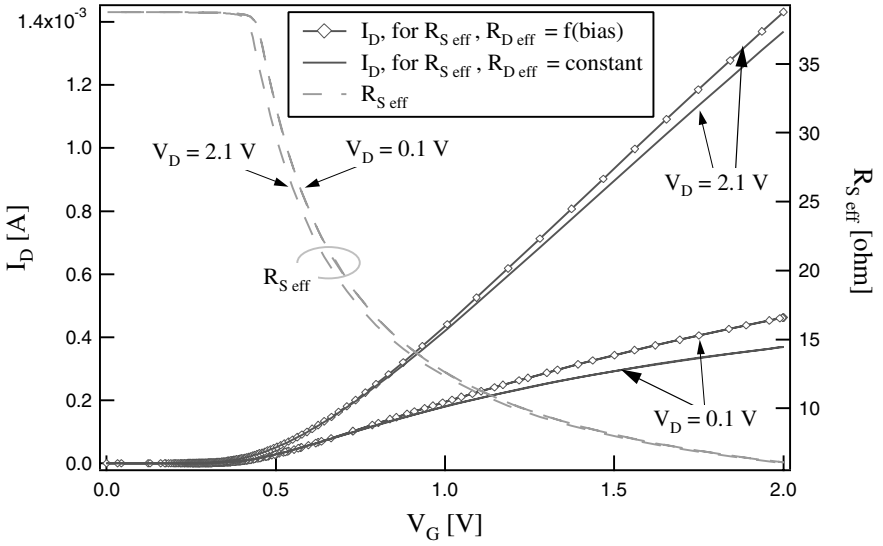


Figure 2. Simulation of I_D and $R_{S\text{eff}}$ versus V_G characteristics for two drain bias.

In Figure 2 and Figure 3 are plotted some results showing the influence of these parasitic elements. All these simulation results in VHDL-AMS are made with a n-channel transistor of $\mathrm{W/L} = 1.5\,\mu\mathrm{m}/0.15\,\mu\mathrm{m}$.

More information of the modeling of these effects and of the simulation results can be found in [18, 19].

4.2.1. Series parasitic resistance

A typical characteristic of series parasitic resistance can be observed in Figure 2, for two different drain bias.

Not taking into account the bias dependent of this resistance introduces some important errors on the drain current level, mainly for small V_D bias. These variations can considerably affect the parameters extraction procedures where, classically, channel length and series resistance are extracted altogether, at small V_D bias [21].

4.2.2. Parasitic capacitances

The dynamic behavior of a MOSFET in deep submicron technology is strongly affected by its extrinsic capacitance formed by the overlap capacitance

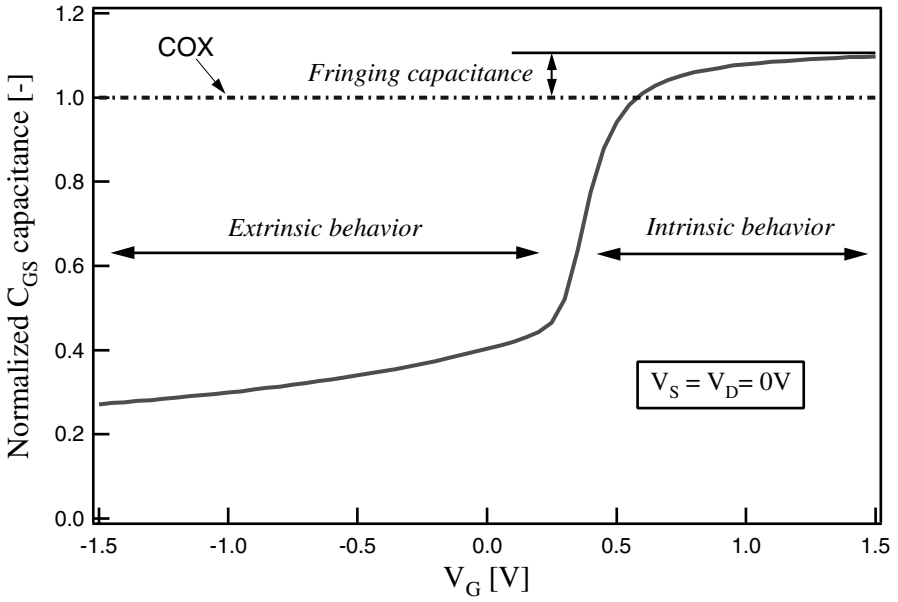


Figure 3. Simulation of the normalized global gate-source capacitance (CGS/COX).

and the fringing capacitance (see Figure 3) [22]. The fringing capacitance is constant, but the overlap capacitance is bias dependent.

As seen in Figure 3, the parasitic capacitances represent a more and more important part of the global capacitance of the MOSFET (more 35% for a 0.15 μm technology) as observed in the accumulation region. The influence of the fringing capacitance can be observed in the strong inversion region; it represents the additional capacitance to COX (COX = W.L.Cox).

4.3. Modeling of an Inverter with Thermo-Electrical Interactions

As the transistor size decreases, thermal interactions between devices on the same chip increase. These thermal effects are constantly amplified by the growing power density, and a failure in their estimation at an early development stage of the design often means extra costs and delays.

For the system designer, one of the major interests of VHDL-AMS is the simplicity with which models involving various physical domains (electronical, thermal, optical, mechanical, etc) can be interconnected. We illustrate this with an example: a CMOS inverter with thermo-electrical interactions.

4.3.1. VHDL-AMS implementation of the EKV model v2.6

In this inverter, the pMOS and nMOS transistor behaviors are described using the EKV MOSFET model v2.6. Several electrical parameters of this model are highly dependent on temperature, namely the threshold voltage, the mobility, the thermal voltage, etc... Their respective temperature variations are taken into account by appropriate coefficients in the model equations [10].

To take the self heating in the MOSFET into account, its packaging must also be considered. Classically, we have modeled the heat diffusion through solid materials by sourcing dissipated power into a thermal RC network [18, 23], which represents the material properties of the different layers. The temperature profile is the result of a heat flow in the thermal network.

Figure 4 shows how thermal-electronic interactions between an n-MOSFET and its direct environment can be modeled.

In such networks, energy conservation states that the sum of all power contributions at a thermal node equals zero, and that the temperature at all terminals connected to a thermal node be identical. Thermal evolution of a system is thus ruled by the very same Kirchhoff laws dictating the behavior of conservative systems: voltage becomes the across quantity temperature and current becomes the through quantity heat flow.

The IEEE standard 1076.1.1-2004 includes the `thermal_system` package that defines the `thermal` nature and its related characteristics. The principle is to introduce a thermal `terminal` and a thermal branch with associated `through` and `across` quantities respectively bound to heat flow (or power) and temperature.

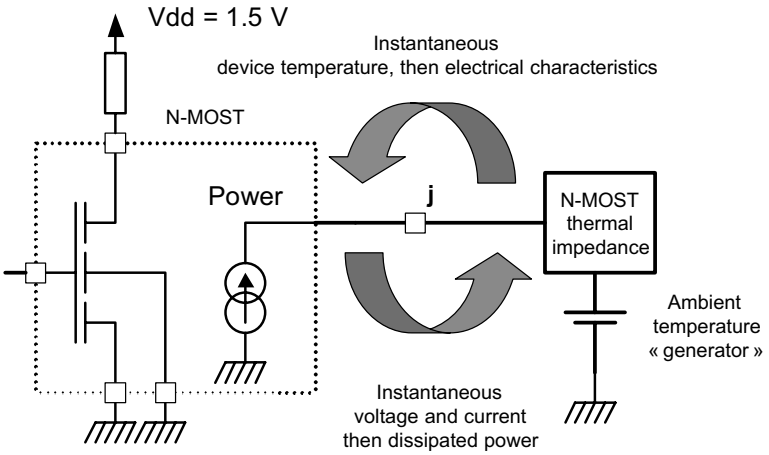


Figure 4. Modeling electro-thermal interactions.

In this paper, we present a simplified version of the EKV MOSFET model with thermo-electrical interactions (Figure 5) [7] as the full version[3] would have needed several pages of code. This single transistor model is valid for both pMOST and nMOST.

```
(1)  library ieee; use ieee.math_real.all;
(2)  library ieee_proposed;
(3)    use ieee_proposed.energy_systems.all;
(4)    use ieee_proposed.electrical_systems.all;
(5)    use ieee_proposed.thermal_systems.all;
(6)  entity mos is
(7)    generic (
(8)      MTYP : real := 1.0; – NMOS: 1.0, PMOS: -1.0
(9)      – geometrical parameters
(10)     WEFF : real := 1.0*MICRO;      – effective channel width
(11)     LEFF : real := 0.15*MICRO;     – effective channel length
(12)     – threshold voltage and substrate body effect parameters
(13)     VT0 : real := 0.4;             – long channel threshold voltage (NMOS!)
(14)     PHI : real := 0.97;            – bulk Fermi potential
(15)     GAMMA: real := 0.71;           – body effect parameter
(16)     – mobility parameters
(17)     KP : real := 453.0*MICRO;      – transconductance parameter
(18)     THETA: real := 50.0*MILLI;     – mobility reduction coefficient
(19)     – temperature coefficients
(20)     TCV : real := 1.5*MILLI;       – temp. coef. of threshold voltage
(21)     BEX : real := -1.5);           – temp. coef. of transcond. parameter
(22)   port (
(23)     terminal td, tg, ts, tb: electrical;
(24)     terminal tj: thermal);
(25) end entity mos;
(26) architecture ekv_simple of mos is
(27)   constant KOQ: real := K/Q;
(28)   constant TEMPREF: real := 300.15;
(29)   – electrical branch quantities
(30)   quantity vg across tg to tb;
(31)   quantity vd across td to tb;
(32)   quantity vs across ts to tb;
(33)   quantity ids through td to ts;
(34)   – thermal branch quantities
(35)   quantity gpower through thermal_ref to tj;
(36)   quantity temp across tj to thermal_ref;
(37)
(38) function i_v (constant v: real) return real is
(39)   variable x: real;
(40) begin
(41)    return (log(1.0 + 0.5*exp(v)))**2;
(42) end function i_v;
```

[3]Different VHDL-AMS simulation results of the thermo-electrical interactions in a MOSFET (with the full version of the EKV v2.6 MOSFET model [20]) can be found in [19].

```
(43)
(44)  function f_id (temp, vg, vs, vd: real) return real is
(45)    variable id, vt, ratio, eg, egref: real;
(46)    variable vto_th, kp_th: real;
(47)    variable vgp_0, vgp, vp, iff, irr, beta, n: real;
(48)  begin
(49)    vt := KOQ*temp + 1.0e-6;
(50)    ratio := abs(temp/TEMPREF + 1.0e-6);
(51)    vto_th := MTYP*(VT0 - TCV*(temp - TEMPREF));
(52)    kp_th := KP*(ratio**BEX);
(53)    vgp_0 := vg - vto_th + PHI + GAMMA*sqrt(PHI);
(54)    vgp :=0.5*(vgp_0+sqrt(vgp_0*vgp_0+1.0e-3));
(55)    vp := vgp - PHI - GMA*(sqrt(vgp+0.25*GMA*GMA)-0.5*GMA);
(56)    iff := i_v((vp - vs)/vt);
(57)    irr := i_v((vp - vd)/vt);
(58)    beta := kp_th*(WEFF/LEFF)*(1.0/(1.0 + THETA*vp));
(59)    n := 1.0;
(60)    return 2.0*n*beta*vt*vt*(iff - irr) + 1.0e-10;
(61)  end function f_id;
(62)
(63)  begin
(64)      ids == MTYP*f_id(temp, MTYP*vg, MTYP*vs, MTYP*vd);
(65)      gpower == abs(ids*(vd - vs));
(66)  end architecture ekv_simple;
```

Figure 5. VHDL-AMS model of a simple EKV MOST model. (n- and p- channel).

As we can see in Figure 5, a traditional VHDL-AMS file first contains references to the used libraries (lines 1–5). In this EKV model, electrical and thermal domains are requested. As previously mentioned in Section 2, we still use a proprietary version of the `thermal_system` package that is available in the `ieee_proposed` library.

For the circuit designer, the most important part is the interface, known as an `entity` in VHDL-AMS (lines 6–25). The interface ports are the four standard electrical pins of a MOSFET (line 23), plus an additional thermal pin to account for dynamic thermal exchanges between the transistor and its environment (line 24). The order in which terminals are specified in a branch quantity declaration defines the direction of the flow.

In VHDL-AMS, the `generic` statement in the `entity` allows the designer to define parameters whose values can be overridden during instantiation of the sub-model. Typically here, the geometrical parameters (`weff` and `leff`)and the electrical parameters (`VTO, PHI, ...`) of the transistor are defined as generic.

The `MTYP` generic parameter allows defining the type of the MOS transistor and also the sign of some relevant voltages and parameters (defined in line 8, used in lines 51 and 64). Note that some actual parameters in the MOS instances must anyway have the right sign (e.g., `VT0` in Figure 5).

The MOSFET behavior is defined in a separate architecture called `EKV_simple` (lines 26–66). Lines 30–36 declare a number of *branch quantities* that correspond in VHDL-AMS to the unknowns of the system of equations to be solved by the analog solver. These branch quantities are defined between two terminals and represent `across` and `through` aspects.

These branch quantities are either `electrical` (lines 30–33) or `thermal` (lines 35–36).

Considering the thermal `port`, the temperature `temp` is measured between the port and the thermal reference (line 36), while the heat `gpower` is flowing out the device from the thermal reference to the port (line 35). This way, the thermal interaction is really bi-directional and the self-heating behavior of the device is properly taken into account with the power computation (line 65).

As the electrical behavior of the EKV MOSFET model is naturally procedural, it is more efficient to use the sequential statements provided in VHDL-AMS. The simultaneous `procedural` statement could be used, but, as it is not yet supported in the Mentor graphics tool, we had to use a function instead, namely the `f_id` function, to implement the computation of the drain current (lines 44 to 61). The equation of the drain current is then implemented in a single simultaneous statement with the appropriate signs for the function arguments to account for the actual model type (line 64). Note that all terminal potentials are defined relatively to the bulk terminal, a specificity of the EKV MOSFET model.

To illustrate the interest of such a model, the simulation of the charge (*QI*) and of the transconductance (*Gm*) versus the gate voltage (*VG*), with and without thermal coupling, is given in Figure 6[4].

As we can see, not taking the thermal coupling into account in a transistor (or a circuit) can lead to some important errors in the estimation of the electrical performances of the device.

4.3.2. VHDL-AMS implementation of the CMOS inverter

The CMOS inverter is composed of one nMOS and one pMOS transistor and is connected to its direct environment as shown in Figure 7.

When located on the same substrate, thermo-electronic interactions take place between the nMOS transistor and pMOS transistor. In this CMOS inverter, the nMOS and the pMOS are thermally interconnected through a coupling thermal resistance: `Rcoupling`. The inverter is excited by a squarewave stimulus. The two thermal networks represent the thermal constants of the various material layers of each transistor.

Figure 8 shows the hierarchical tree of sub-models in the VHDL-AMS testbench.

[4]These results are obtained with the full VHDL-AMS code [20].

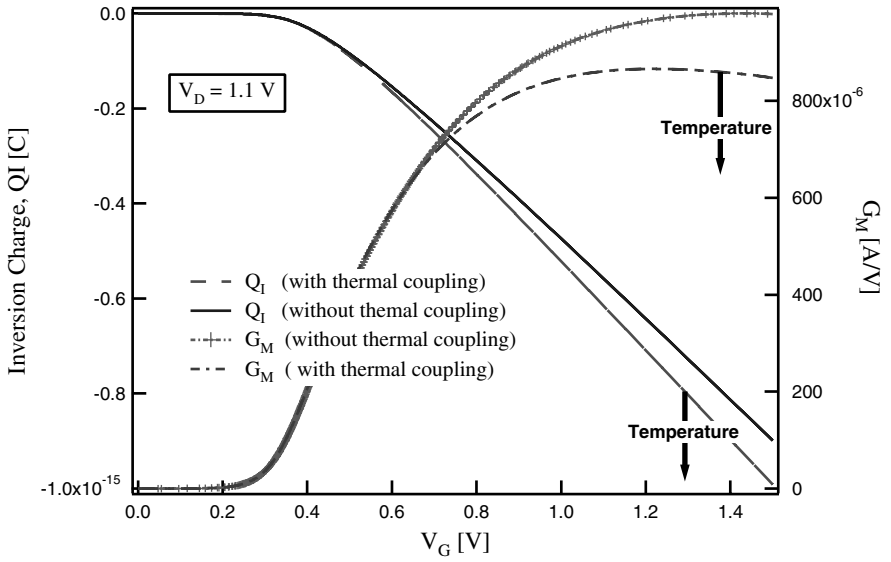


Figure 6. Characteristics of the inversion charge *QI* and the transconductance *Gm* vs. *VG*.

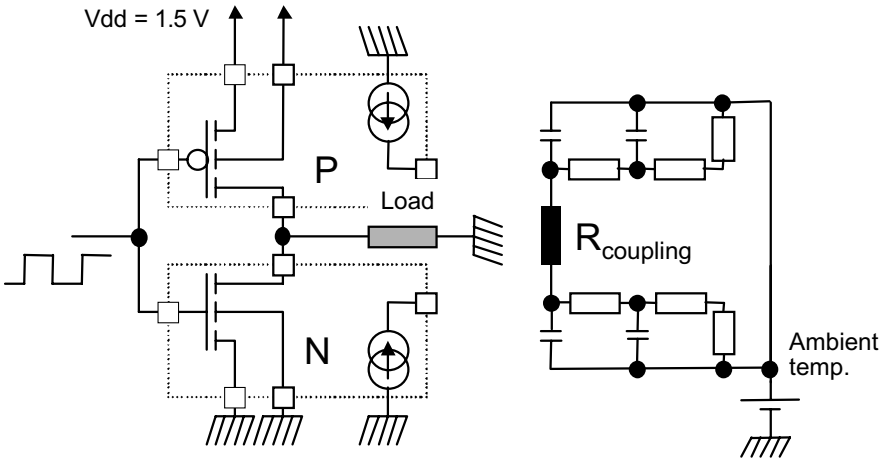


Figure 7. The CMOS inverter and its direct environment.

The thermal network is modeled by thermal resistor and thermal capacitors. The VHDL-AMS models of the thermal capacitance, the thermal resistance, and the ambient heat source (thermal generator) are given in Figure 9 [7, 24, 25]. The thermal resistor and capacitor models are straightforward equivalents of their electrical counterparts.

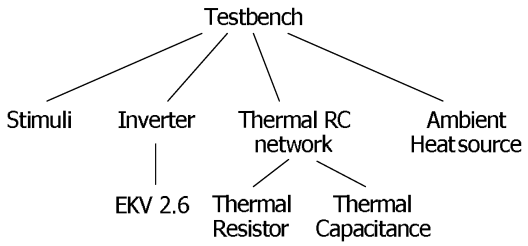


Figure 8. The VHDL-AMS testbench hierarchical tree.

```
(1)   – Behavioural model of a thermal capacitor
(2)   library ieee_proposed;
(3)      use ieee_proposed.energy_systems.all;
(4)      use ieee_proposed.thermal_systems.all;
(5)
(6)   entity capth is
(7)      generic (CVAL: real := 0.1*PICO);
(8)      port (terminal tp, tm: thermal);
(9)   end entity capth;
(10)
(11)  architecture bce of capth is
(12)        quantity temp across hfl through tp to tm;
(13)  begin
(14)        hfl == CVAL*temp'dot;
(15)  end architecture bce;
(16)
(17)  – Behavioural model of a thermal resistor
(18)  entity resth is
(19)        generic (RVAL: real := 1.0*KILO);
(20)        port (terminal tp, tm: thermal);
(21)  end entity resth;
(22)
(23)  architecture bce of resth is
(24)        quantity temp across hfl through tp to tm;
(25)  begin
(26)        temp == RVAL*hfl;
(27)  end architecture bce;
(28)
(29)  – Behavioural model of a thermal generator
(30)  entity genetherm is
(31)        generic ( tambient : real := 300.0);
(32)        port (terminal tlp : thermal);
(33)  end;
(34)
(35)  architecture equ of genetherm is
(36)        quantity temp across power through tlp to thermal_ground;
(37)  begin
(38)        temp == tambient ;
(39)  end;
```

Figure 9. VHDL-AMS models of the thermal components.

```
(1)   library ieee_proposed;
(2)     use ieee_proposed.energy_systems.all;
(3)     use ieee_proposed.electrical_systems.all;
(4)     use ieee_proposed.thermal_systems.all;
(5)   entity cmos_inv is
(6)     generic (
(7)       WN: real := 15.0*MICRO; – NMOS channel width
(8)       LN: real := 0.15*MICRO;  – NMOS channel length
(9)       WP: real := 15.0*MICRO; – PMOS channel width
(10)      LP: real := 0.15*MICRO); – PMOS channel length
(11)    port (
(12)      terminal tin, tout, tvdd, tvss: electrical;
(13)      terminal tjn, tjp: thermal);
(14)  end entity cmos_inv;
(15)
(16)  architecture str of cmos_inv is
(17)  begin
(18)    PMOS: entity work.mos(ekv_simple)
(19)      generic map (
(20)        MTYP => -1.0, WEFF => WP, LEFF => LP,
(21)        VT0 => -0.4, TCV => -1.5*MILLI)
(22)      port map (
(23)        td => tout, tg => tin, ts => tvdd, tb => tvdd, tj => tjp);
(24)    NMOS: entity work.mos(ekv_simple)
(25)      generic map (
(26)        MTYP => 1.0, WEFF => WN, LEFF => LN,
(27)        VT0 => 0.4, TCV => 1.5*MILLI)
(28)      port map (
(29)        td => tout, tg => tin, ts => tvss, tb => tvss, tj => tjn);
(30)  end architecture str;
```

Figure 10. VHDL-AMS structural model of the CMOS inverter.

The CMOS inverter is a structural model that instantiates two components: one pMOS transistor called `PMOS` (Figure 10, lines 18 to 23) and one nMOS transistor called `NMOS` (Figure 10, lines 24 to 29). Both the generic parameters and the port associations use the named association mechanism for improved readability.

Figure 11 shows the temperature evolution in the inverter for two different values of `Rcoupling`. As expected, for a small value of `Rcoupling`, the temperature in the N and P transistors are tightly linked (curves 1 and 2). For a higher value, Figure 11 shows the free temperature evolution of each transistor (curves 3 and 4).

For simulation purpose, the values of the capacitances and resistances of the thermal network have voluntary been overstated to emphasize the thermal effects.

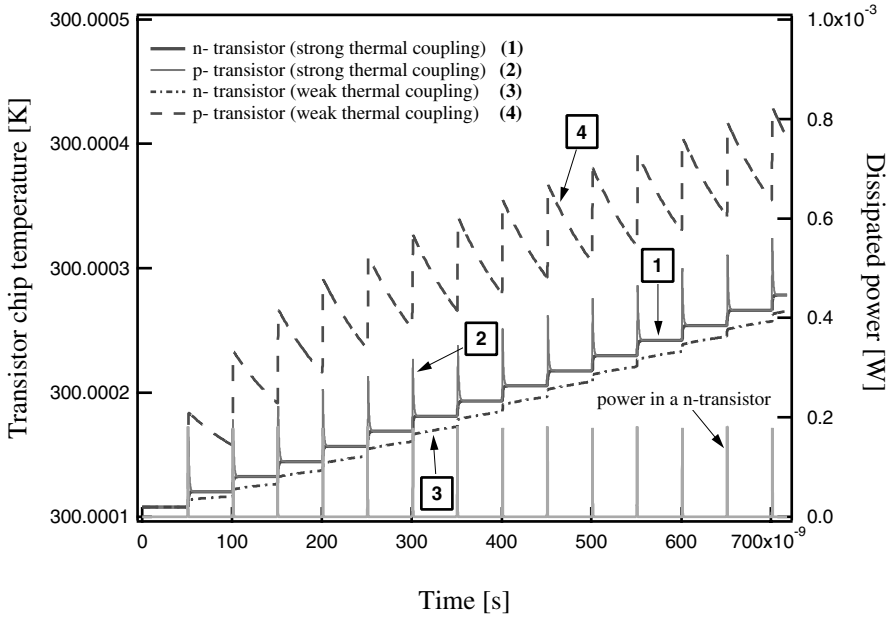


Figure 11. Simulation of the n/p – channel transistor chip temperature variation versus time during commutation (in an inverter), for two values of thermal coupling.

5. Accounting for the Quantum Effects in an Advanced MOSFET Model

As CMOS technology tends towards ever thinner gate oxide and higher substrate doping concentration, the quantum effects are more and more significant. From a physical point of view, they result in a change in the relationship between charges and applied voltages. In previous works, we have shown that in the context of a surface potential model, we only need to compute the exact value (i.e. the quantum value) of the surface potential to get a coherent model [26, 27]. The proposed model fully accounts for the quantum effects and is able to accurately describe all major characteristics of MOSFET. It does not require either definition of an effective oxide thickness or use of additional parameters.

Based on the core of the MM11 model [28], we have developed new concepts to compute the exact value of the surface potential, i.e. accounting for the quantum effects. The model covers all operating regions from accumulation to inversion and is valid for all bias conditions. This section is organized as follows. First, we describe the physical basis of our analytical and quantum surface potential model. The straightforward use of a charge sheet model

(drift-diffusion approximation) is then discussed. Next, we detail the VHDL-AMS implementation of the quantum model for an n-MOS transistor. Finally, the simulation results obtained with the VHDL-AMS model are presented.

5.1. Modeling the Quantum Mechanical Effects

In advanced CMOS technologies, the use of thin gate oxides and high substrate doping levels results in a very high normal field at the Si–SiO_2 interface, so that the energy spectrum consists of a set of discrete energy levels, where the first allowed energy level does no longer coincide with the bottom (top) of the conduction (valence) band. Figure 12 shows the energy band diagram of an n-MOS transistor biased in strong inversion. It appears that the quantum mechanical effects (QME) increase the apparent bandgap of the semiconductor $(\Delta E_g = E_0 - E_c)$. The same reasoning is valid for the accumulation region as well. Within the context of surface-potential-based models, we have demonstrated that the quantum effects can be fully taken into account by the new concept of pseudo bandgap widening. The reader is referred to references [26, 27, 29] for full details of the procedure.

Once the explicit relationship between the quantum increment/decrement of the surface potential and the gate and source/drain voltages is known, incorporating this relationship into the core of the MM11 model (i.e. a classical

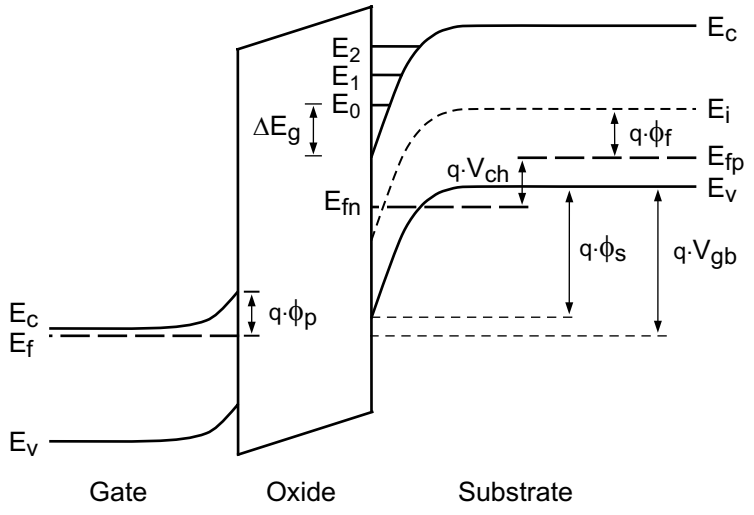


Figure 12. Energy band diagram of an n-channel MOSFET (strong inversion mode). ϕ_s is the surface potential, ϕ_p is the band bending in the gate due to the polydepletion effect, V_{ch} is the channel potential (electron quasi-Fermi potential) and ϕ_f the intrinsic Fermi potential.

description of the surface potential without QME) [28] allows us to obtain an accurate analytical and quantum surface potential model valid from accumulation to strong inversion, and from linear to saturation region [29].

Finally, taking into account the quantum effects does not make the new model more computational demanding and does not introduce any additional parameter with respect to a classical description of the surface potential. Figure 13 shows a comparison between the surface potential computed with the new model and the results obtained by a self-consistent resolution of the Schrödinger and Poisson equations.

A major interest of a surface potential model is that it enables a straightforward use of a charge sheet model since all charges in the latter explicitly depend on the surface potential value [30]. As our model computes the surface potential analytically, the use of the drift-diffusion approximation does not require time consuming iterations to calculate the surface potential (at the source and drain ends). This means that the major advantage of common piece-wise models does no longer hold, and consequently the surface-potential-based MOSFET models are the best candidates to be chosen as new standard compact MOSFET models.

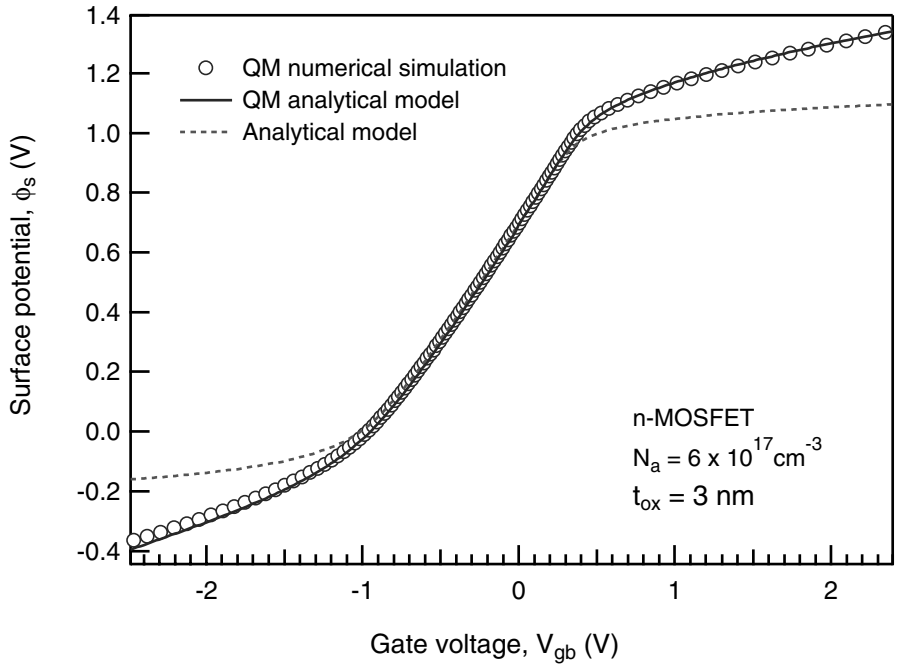


Figure 13. Surface potential analytically computed as a function of gate voltage. The channel potential V_{ch} is set to zero in this simulation.

5.2. VHDL-AMS Implementation

In Figure 14, we present the skeleton of the VHDL-AMS code for the quantum surface potential model (the full code is available on a website [20]). The VHDL-AMS code first contains references libraries needed to parse the model (lines 1–2). For the model end-user (circuit designer), the most important part of the model is the interface, contained in what is called the VHDL-AMS `entity` (lines 33–39). The model interface is decomposed into the specification of generic parameters (lines 34–37) and of the interface ports (line 38). The `generic` statement allows the designer to define its own values for the model parameters. Typically, geometrical W and L transistor parameters are defined as generic. The `mosfet` entity contains four terminals (G, D, S and B stand for the gate, drain, source and bulk terminal respectively), all of electrical type. All the terminals are part of a `port` statement.

```
(1)   library ieee;
(2)     use ieee.electrical_systems.all; use ieee.math_real.all;
(3)
(4)   – Functions declaration
(5)   package mm11_functions is
(6)     pure function phis1_qm_pd(Cox,Vg,…,PDE:real) return real; –acc.
(7)     pure function phis2_qm_pd(Cox,Vg,Vch,…,PDE:real) return real; –inv.
(8)     …/…
(9)   end;
(10)  – Functions definitions
(11)  package body mm11_functions is
(12)  – Quantum description of the surface potential (accumulation)
(13)    pure function phis1_qm_pd(Cox,Vg,…,PDE:real) return real is
(14)      variable ret :real;
(15)    begin
(16)      ret := …; return ret;
(17)    end phis1_qm_pd;
(18)  – Quantum description of the surface potential (inversion)
(19)    pure function phis2_qm_pd(Cox,Vg,Vch,…,PDE:real) return real is
(20)      variable ret :real;
(21)    begin
(22)      ret := f_qm_pd(Cox,Vg,Vch,…,PDE) +
(23)      phit*log(((((2.0/gamma)*(Vg-Vfb-
(24)      psiB(Cox,Vg,Vch,…,PDE)))
(25)      / (1.0+sqrt(1.0+(4.0/gamma_p**2)*
(26)      (Vg-Vfb-psiB(Cox,Vg,Vch,…,PDE)))))**2 -
(27)      f_qm_pd(Cox,Vg,Vch,…,PDE)+phit) / phit);
(28)      return ret;
(29)    end phis2_qm_pd;
(30)      …/…
(31)  end mm11_functions;
(32)
(33)  entity mosfet is
(34)    generic(W :real := 1.0e-6; – Gate width
```

```
(35)       L :real := 1.0e-6; – Gate length
(36)       Na :real := 5.0e23; – Substrate doping level
(37)       …/…);
(38)     port (terminal G,D,S,B :electrical);
(39)   end entity mosfet;
(40)
(41)   architecture quantum_polydep of mosfet is
(42)     constant T :real := 300.0; …/…
(43)     quantity Qg1,Qg2,Qg,Cgg1,Cgg2,Cgg :real;
(44)     quantity Idiff,Idrift :real; …/…
(45)     quantity Ids through D to S;
(46)     quantity Vdb across D to B; quantity Vsb across S to B;
(47)     quantity Vgb across G to B;
(48)   begin
(49)     …/…
(50)   – Gate charge & gate transcapacitance
(51)     Qg1 == W*L*Cox*(Vgb-Vfb-phip-phis1_qm_pd(Cox,Vgb,…,PDE));
(52)     Qg2 == 0.4*W*L*Cox*(Vgb-Vfb-phip-phis2_qm_pd(Cox,Vgb,Vdb,…,PDE))
(53)              +0.6*W*L*Cox*(Vgb-Vfb-phip-phis2_qm_pd(Cox,Vgb,Vsb,…,PDE));
(54)     Cgg1 == Qg1'dot;
(55)     Cgg2 == Qg2'dot;
(56)       if Vgb'above(0.0) use
(57)       Qg == Qg2; Cgg == Cgg2/(W*L*Cox);
(58)       else
(59)       Qg == Qg1; Cgg == Cgg1/(W*L*Cox);
(60)       end use;
(61)   – Drain current
(62)     Idrift == …; Idiff == …; Ids == Idrift + Idiff;
(63)     …/…
(64)   end architecture quantum_polydep;
```

Figure 14. Skeleton of the VHDL-AMS code for the modified MM11 model.

The MOSFET behavior is defined in a separate architecture called `quantum_polydep` (lines 41–64). Lines 43–47 declare *quantities*.

In the lines 45–47, a number of *branch quantities* are declared. A number of so-called *free quantities* (i.e. quantities not bound to any terminal) are also declared (lines 43–44). These quantities are mainly used to break down complex relationships into more manageable and understandable pieces.

The electrical behavior of the surface potential model is actually procedural so it is more efficient to use the sequential statements proposed by VHDL-AMS. However, as the simultaneous `procedural` statement is not yet supported in the Simplorer tool, we use different functions instead. All the needed functions are defined in a package called `mm11_functions` (lines 5–31). This package is divided into two units. The first unit includes the declaration of the functions prototypes (lines 5–9) while the second unit includes the functions bodies (lines 11–31).

For instance, the `phis2_qm_pd` function given in lines 19–29 corresponds to the following mathematical relationship:

$$\phi_{s2\_qm\_pd} = f_{\_qm\_pd} + \phi_t \cdot \ln\left\{\left[\left(\frac{\frac{2}{\gamma}\cdot(V_g - V_{fb} - \psi_B)}{1+\sqrt{1+\frac{4}{\gamma_p^2}\cdot(V_g - V_{fb} - \psi_B)}}\right)^2 - f_{\_qm\_pd} + \phi_t\right]/\phi_t\right\}$$

Next, making use of the functions, we can easily implement useful quantities such as, for example, the gate charge/transcapacitance (lines 51–60) and the drain current (lines 62). In our model, all the free quantities (simultaneous statements) are functions of the surface potential. They just depend on two functions, namely `phis1_qm_pd` and `phis2_qm_pd`. For example, lines 50–60 detail the implementation of the gate transcapacitance. The simultaneous `if` statement has been used to select the `phis1_qm_pd` or `phis2_qm_pd` function which corresponds to the formulation of the surface potential in accumulation or inversion. Note the use of the `'dot` attribute to denote a first-order time derivative of the quantity prefix.

5.3. Simulation Results

The VHDL-AMS model previously discussed has been implemented using Simplorer 6.0 from Ansoft. Four different architectures for the `mosfet` entity have been implemented. They allow the user to choose between classical or quantum surface potential models and give the possibility to take into account or not the polydepletion effect (PDE). In fact, only the architecture called `quantum_polydep` (see Figure 14) is useful since it describes an n-MOS transistor using the full model (QME+PDE). The three others have been written for comparison purpose.

We have tested the different n-MOSFET architectures by applying a 1V/s ramp on the gate terminal in a transient simulation. Both source and bulk terminals of the device are connected to the ground and the drain-to-bulk voltages are set to 0.1V. Figure 15 shows the simulated gate transcapacitance ($C_{gg} = dQ_g/dV_{gb}$) for different architectures of the `mosfet` entity.

Since the derivative over time of the gate-to-bulk voltage equals one, the quantity `Qg'dot` (lines 54–55 in Figure 14) is simply equal to the gate transcapacitance ($dQ_g/dt = C_{gg} \times dV_{gb}/dt = C_{gg}$).

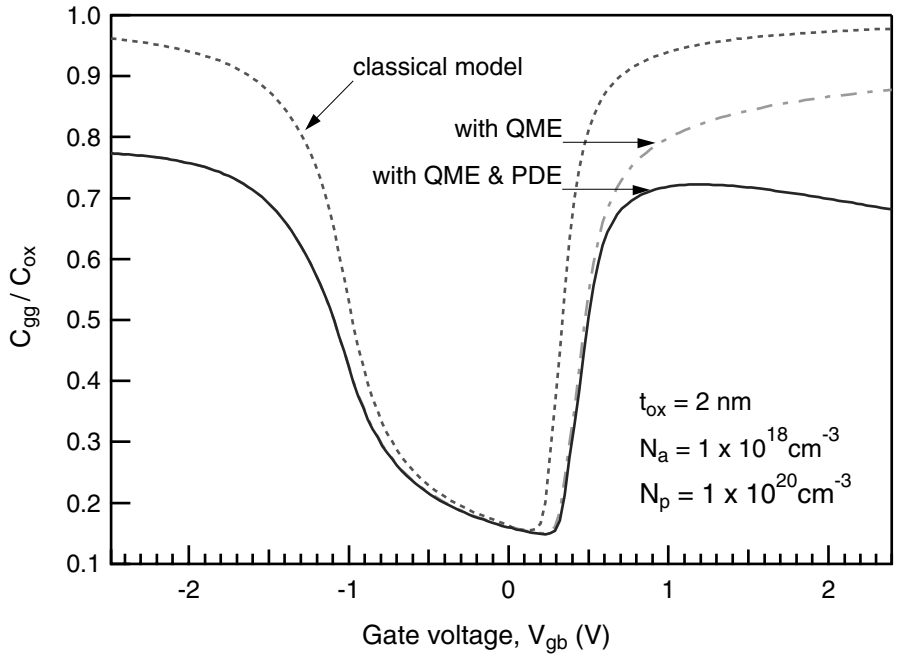


Figure 15. Normalized gate transcapacitance as a function of gate voltage for different architectures (classical, quantum = QME; quantum with polydepletion effect = QME + PDE).

Figure 16 gives the *I–V* simulation extracted from the same transient simulation. The drain current I_{ds} is obtained as the sum of the drift and diffusion currents and exhibits an excellent behavior from weak to strong inversion.

6. Conclusions

The presented case studies show that VHDL-AMS can be successfully used to implement low-level models, such as EKV 2.6 and MM11 models of MOSFET devices. The main point is that the physical equations of the models can actually be written "as is" in the model source code. The only limitations in this straightforward translation do not come from the language itself, but from the available simulation tools, Advance-MS from Mentor Graphics, and Simplorer from Ansoft. The lack of support in these commercial tools for a procedural statement forces us to use functions with numerous parameters to avoid a whole set of simultaneous statements.

Taking the various domains involved in the modeling task is not complicated either. We have shown that the basic models can be easily enhanced to include major physical effects like self-heating, extrinsic aspects and quantum

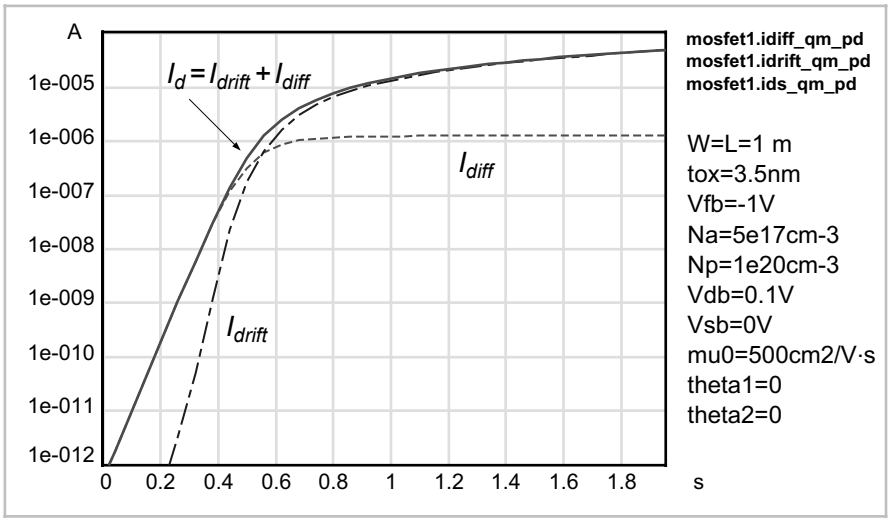


Figure 16. Screen dump of a VHDL-AMS simulation showing both drift and diffusion components of the drain current for an n-channel MOSFET.

effects, as the VHDL-AMS language naturally supports multi-domain (thermal, mechanical, fluidic, etc) modeling.

In a near future, we plan to use these compact multi-domain models to design analog circuits like an operational amplifier, and digital subsystems like a nand gate and flip-flop, that take into account thermo-electrical interactions. We also plan to use the bond-graph modeling approach experienced in [31] to propose accurate models of isFET (ion sensitive FET), that could eventually lead to the development of biosensor simulation models. Some models of advanced devices such as DG-MOSFET are also under development [32].

Acknowledgements

The authors would like to thank Dr. Wladek Grabinski for his strong implication in MOS-AK activities (see MOS-AK web site: http://www.mos-ak.org/).

References

[1] 1076.1-1999 IEEE Standard VHDL Analog and Mixed-Signal Extensions Language Reference Manual, *IEEE Press, ISBN 0-7381-1640-8*, **1999**.

[2] 1076.1.1-2004 IEEE Standard VHDL Analog and Mixed-Signal Extensions – Packages for Multiple Energy Domain Support, *IEEE Press, ISBN 0-7381-4646-5*, **2005**.

[3] Ashenden, P.; Peterson, G.D.; Teegarden, D.A. "The System Designer's Guide to VHDL-AMS", Morgan Kaufman Publishers, **2002**, *ISBN 1-55860-749-8*.
[4] Christen, E.; Bakalar, K. "VHDL-AMS-a hardware description language for analog and mixed-signal applications", *IEEE Trans. Circ. & Syst. - Part II*, **1999**, *46(10)*, 1263–1272.
[5] Frey, P.; O'Riordan, D. "Verilog-AMS: Mixed-signal simulation and cross domain connect modules," *Proc. of the IEEE/ACM Int. Workshop Behav. Modeling Simul. (BMAS)*, **2000**, 103–108.
[6] http://www.eda.org/verilog-ams/
[7] Pêcheux, F.; Lallement, C.; Vachoux, A. "VHDL-AMS and Verilog-AMS as Alternative Hardware Description Languages for Efficient Modeling of Multi-Discipline Systems", *IEEE Trans. Comput.–Aided Design Integrated Circ. and Syst.*, **2005**, *24*, 204–225.
[8] BSIM 4.2 Manual. http://www-device.eecs.berkeley.edu/∼bsim (**April 2001**).
[9] The MM9 model. http://www.semiconductors.philips.com/philips_models/mos models (**2001**).
[10] Bucher, M.; Lallement, C.; Enz, C.C.; Théodoloz, F.; Krummenacher, F. "The EPFL-EKV MOSFET Model, Version 2.6", *Technical Report*, **1999**, [Online] http://legwww.epfl.ch/ekv/
[11] Foty, D. "MOSFET modelling with Spice". *Principles and Practice*, Englewood Cliffs, NJ: Prentice-Hall, *ISBN: 0-13-227935-5*, **1997**.
[12] Tsividis, Y.P. "Operation and Modelling of the MOS Transistor", second edition, New York: McGraw-Hill Book Company, *ISBN: 0070655235*, **1999**.
[13] Bucher, M.; Enz, C.C.; Krummenacher, F.; Sallese, J.-M.; Lallement, C.; Porret, A.-S. "The EKV 3.0 compact MOS transistor model: accounting for deep-submicron aspects", *Proc. MSM Int. Conf., Nanotech 2002*, **2002**, 670–673.
[14] Miura-Mattausch, M.; Mattausch, H.J.; Arora, N.D.; Yang, C.Y. "MOSFET modelling gets physical", *IEEE Circ. and Dev.*, **2001**, *17(6)*, 29–36.
[15] van Langevelde, R.; Scholten, A.J.; Havens, R.J. "Advanced compact MOS modelling", *Proc. ESSDERC*, **2001**, 81–90.
[16] Chen, T.L.; Gildenblat, G. "Analytical approximation for the MOSFET surface potential", *Solid-State Electron.*, **2001**, *45*, 335–339.
[17] Gildenblat, G.; Li, X.; Wang, H.; Lu, W.; van Langevelde, R.; Scholten, A.J.; Smit, G.D.J.; Klaassen, D.B.M. "Introduction to PSP MOSFET Model", *Proc. the MSM 2005 Int. Conf., Nanotech 2005*, **2005**.
[18] Lallement, C.; Pêcheux, F.; Hervé, Y. "VHDL-AMS Design of a MOST Model Including Deep Submicron and Thermal-Electronic Effects", *Proc. IEEE/ACM Int. Workshop Behav. Modeling Simul. (BMAS)*, **2001**, *91–96*.
[19] Lallement, C.; Pêcheux, F.; Hervé, Y. "A VHDL-AMS Case study: The incremental Design of an Efficient 3rd generation MOS Model of Deep Sub Micron Transistor", In *SOC Design Methodologies*, M. Robert, B. Rouzeyre, C. Piguet and M.-L. Flottes, Eds., Boston, Hardbound, Kluwer Academic Publishers, ISBN 1-4020-7148-5, **July 2002**, 349–360.
[20] http://lsmwww.epfl.ch/models/compact/
[21] Arora, N. "Mosfet Models for VLSI Circuit Simulation: Theory and Practice", *Computational Microelectronics*, Wien, New York: Springer Verlag, **1993**.
[22] Prégaldiny, F.; Lallement, C.; Mathiot, D. "A simple efficient model of parasitic capacitances of deep-submicron LDD MOSFETs", *Solid-State Electron.*, **2002**, *46*, 2191–2198.
[23] Lallement, C.; Bouchakour, R.; Maurel, T. "One-dimensional Analytical Modeling of the VDMOS Transistor Taking Into Account the Thermoelectrical Interactions", *IEEE Trans. Circ. & Sys., Part. I*, **1997**, *44(2)*, 103–111.

[24] Lallement, C.; Pêcheux, F.; Grabinski, W. "High Level Description of Thermodynamical effects in the EKV 2.6 MOST Model", *9th Int. Conf. Mixed Design of Integrated Circ. & Sys. (MIXDES 2002)*, Wroclaw/Poland, **June 2002**, 45–50.

[25] http://lsmwww.epfl.ch/tcad_paper/.

[26] Prégaldiny, F.; Lallement, C.; van Langevelde, R.; Mathiot, D. "An advanced explicit surface potential model physically accounting for the quantization effects in deep-submicron MOSFETs", *Solid-State Electron.*, **2004**, *48(3)*, 427–435.

[27] Prégaldiny, F.; Lallement, C.; Mathiot, D. "Accounting for quantum mechanical effects from accumulation to inversion, in a fully analytical surface-potential-based MOSFET model", *Solid-State Electron.*, **2004**, *48(5)*, 781–787.

[28] van Langevelde, R.; Klaassen, F.M. "An explicit surface-potential-based MOSFET model for circuit simulation", *Solid-State Electronics*, **2000**, *44*, 409–418.

[29] Prégaldiny, F.; Lallement, C. "Fourth generation MOSFET model and its VHDL-AMS implementation", *Int. J. Numerical Modelling: Electronic Networks, Dev. and Fields*, **2005**, *18*, 39–48.

[30] Brews, J.R. "A charge sheet model of the MOSFET", *Solid-State Electron.*, **1978**, *21*, 345–355.

[31] Pêcheux, F.; Allard, B.; Lallement, C.; Vachoux, A.; Morel, H. "Modeling and simulation using bond graphs and VHDL-AMS", *Int. Conf. Bond Graph Modeling & Simul. (ICBM'2005)*, New Orleans - USA, **2005**, 149–155.

[32] Prégaldiny, F.; Krummenacher, F.; Diagne, B.; Pêcheux, F.; Sallese, J.-M.; Lallement, C. "Explicit modeling of the double-gate MOSFET with VHDL-AMS", *submitted to Int. J. Numerical Modelling: Electronic Networks, Dev. and Fields*, **July 2005**.

Chapter 10

COMPACT MODELING IN VERILOG-A

Boris Troyanovsky, Patrick O'Halloran, and Marek Mierzwinski
Tiburon Design Automation, Inc.
E-mail: boris@tiburon-da.com

Abstract: Historically, compact transistor models have been developed using general-purpose programming languages such as C or Fortran, with the resulting source code specifically targeted to a given circuit simulator's proprietary model interface. Although this approach has allowed for the creation of robust and efficient compact models, it has nevertheless resulted in a situation where the model development process is lengthy, the models are not portable across the various simulation environments, and where the model development facilities are often not open to independent model developers. The advent of analog hardware description languages (AHDLs) over the last several years promises to address the aforementioned issues by providing a portable, robust, and efficient platform for analog model development. In this chapter, we describe the Verilog-A language and explore the numerous benefits it provides in the area of compact modeling.

Key words: Verilog-A; AHDL; compact device model; transistor model; analog model.

1. Introduction and Overview

The availability of accurate, robust, and efficient compact models is critical to the successful utilization of any circuit simulation tool. As new physical effects manifest themselves due to shrinking geometries, and as an increasingly wide variety of highly specialized device technologies (e.g., RF CMOS, SiGe, III–V) become available to analog circuit designers, the need for rapid development and distribution of advanced semiconductor device

W. Grabinski, B. Nauwelaers and D. Schreurs (eds.),
Transistor Level Modeling for Analog/RF IC Design, 271–291.

models becomes more acute than ever. Traditionally, circuit simulators have relied largely on "built-in" semiconductor device models. Such built-in devices – typically implemented using general-purpose programming languages like C, C++, or Fortran – are targeted specifically to the interface and internal data structures of their host simulator, and are thus inherently non-portable. Facilities for adding custom models (or "user-defined devices") have been made available in some simulation environments, but such interfaces have typically been non-standard, non-portable, and inefficient. New model creation under these conditions was thus a time-consuming and error-prone endeavor.

The rapidly increasing availability and adoption of analog HDLs such as Verilog-A [1, 2] offers the promise of a comprehensive solution to the aforementioned analog model development and deployment problem. Initially conceived as a general-purpose analog modeling language, Verilog-A has over the past several years become increasingly viewed as a leading candidate for new compact model development [3–6]. Although the language has always been applicable across the full range of analog modeling tasks – from behavioral event-driven models all the way down to the transistor level – early Verilog-A implementations were interpreted solutions, and were not viewed as being viable alternatives to hand-coded built-in device models. The recent rise in interest for Verilog-A based compact model development has resulted in compiled solutions becoming available, with an ongoing emphasis on improved simulation performance.

The use of standardized, special-purpose analog HDLs such as Verilog-A allows device modeling experts to focus on their area of expertise, rather than on the underlying simulator-specific implementation details. The increased level of abstraction means that the model developer can focus on model behavior, and let the underlying implementation automatically take care of mundane (and often simulator-specific) details such as matrix stamping and loading, analysis-specific data structures, symbolic derivative computation, and so forth. The device modeling engineer is thus shielded from the idiosyncrasies and complexities of the various device interfaces in existence today.

2. Verilog-A Language Fundamentals

For model developers accustomed to working in a standard programming language such as C or Fortran, the switch to Verilog-A syntax should be straightforward and painless. The language is relatively succinct and compact, and is well-suited to analog model development. Several academic and industrial model development groups now use Verilog-A as a key part of their development methodology.

2.1. Introduction by Example

To illustrate the straightforward and intuitive nature of Verilog-A source code, we consider the following simple example.

```
module simple_diode(pos, neg);
inout pos, neg;
electrical pos, neg;

parameter real Area = 1.0 from (0:inf);
parameter real Is=1e-14 from [0:inf);
parameter real n = 2 from (0:inf);
parameter real Cjo=0 from [0:inf);
parameter real Phi = 0.7, m = 0.5, tt = 1p;

real Id, Qd;

analog begin
   Id = Area*Is*(limexp(V(pos, neg)/(n*$vt))-1);
   Qd = tt*Id + Area*V(pos, neg)*Cjo/
        pow((1-V(pos, neg)/Phi), m);
   I(pos, neg) <+ Id + ddt(Qd);
end
endmodule
```

The fundamental structural unit within Verilog-A is the module. In the first line of the code fragment above, we see that the module is named "simple_diode", and that it has two terminal connections (or ports, in Verilog-A parlance). The language supports the presence of non-electrical domains, such as electromechanical or thermal; for this simple diode example, the terminals (*pos* and *neg*) are labeled as being "electrical". (Some compact models incorporate thermal effects, and would thus use a "thermal" discipline for the thermal node.) Internal nodes are declared using the same syntax: if a discipline declaration is present for a node whose name does not match the module port list, that node becomes an internal node within the enclosing module.

The diode's parameters, including default values and allowable ranges, are specified after the terminal disciplines, and two local real variables (*Id* and *Qd*) are then declared. Both real and integer-valued quantities are allowed, and arrays of variables (or parameters) are allowed as well. In a subsequent Section (2.7) we will also encounter the Verilog-A specific variable type known as a "genvar".

Following the variable and parameter declarations, we come to the heart of the model's numerical description – the analog block. Each module can contain at most one analog block, where the module's analog behavior is specified. Because Verilog-A allows hierarchical constructs (Section 2.8), some modules can merely instantiate other modules as child instances and connect them electrically. In these cases, the analog block need not be present.

Our simple diode example has no hierarchical constructs within it, and the diode description resides solely within the analog section. Straightforward mathematical expressions are used to assign physically meaningful values to *Id* (the diode current) and *Qd* (the charge). The resulting current is then directed to the output terminals via the contribution statement:

```
I(pos, neg) <+ Id + ddt(Qd);
```

So long as the target of the contribution statement does not switch from current to voltage (or vice versa), the contributions are all additive. The following two statements would be equivalent to the previous one:

```
I(pos, neg) <+ Id;
I(pos, neg) <+ ddt(Qd);
```

In the next several sections, we present a more detailed overview of the Verilog-A language structure, with particular emphasis on constructs important to the compact model developer.

2.2. Contributions and Branches

Verilog-A uses the so-called "source/probe" formulation for describing the behavior of electrical networks. Consider a pair of electrical nodes, named $n1$ and $n2$. As we saw in the previous section, we can "probe" the voltage between them via the expression $V(n1, n2)$. To insert a current source (a "flow-branch" or "current branch" in Verilog-A parlance) between the two nodes, we would use the contribution statement:

```
I(n1, n2) <+ Idc;
```

Similarly, to insert a voltage source (also called a "voltage branch" or a "potential source") between nodes $n1$ and $n2$, the contribution statement:

```
V(n1, n2) <+ Vdc;
```

would be used. The presence of either of these two contribution statements introduces an "unnamed branch" between the two nodes. Explicit named branches can also be introduced via declarations of the form:

```
branch (n1, n2) br_res;
branch (n1, n2) br_cap;
branch (n1, n2) br_ind;
```

and contributed to by statements such as:

```
I(br_res) <+ V(br_res)/R;
I(br_cap) <+ C*ddt(V(br_cap));
```

```
V(br_ind) <+ L*ddt(I(br_ind)) + RL*I(br_ind);
```

Explicitly named branches can be useful in those cases where the user is interested in current flow through the named branch only, either for direct output or for use in another expression:

```
Pdiss_L  = I(br_ind)*I(br_ind)*RL;
Pdiss_R1 = I(br_res)*I(br_res)*R;
```

As we explain in more detail later, most compact modeling applications should attempt to probe *voltage* (i.e., use $V(\ldots)$ only on the right hand side) and contribute to *current* (i.e., use $I(\ldots)$ only on the left hand side) whenever possible. Failure to do so may result in the introduction of additional state variables, causing the simulation to be slower and more memory-intensive than would otherwise be the case. For example, a nonlinear capacitor should be implemented as:

```
I(p, n) <+ ddt(cap(V(p, n));
```

rather than the alternate (and usually less efficient) choice:

```
V(p, n) <+ idt(f(I(p, n));
```

In most implementations, the second choice will result in the introduction of additional state variables into the system.

For some components, of course, the preceding rule of thumb is not applicable. A truly voltage-controlled component such as an inductor should be implemented as:

```
V(p, n) <+ ddt(phi(I(p, n));
```

where (for the sake of generality) we have used the analog function "phi" to refer to the potentially nonlinear inductance characteristic. Although this formulation will introduce an extra state variable into the system, the voltage-controlled nature of the component makes this intrinsically necessary. The alternate integral-based implementation:

```
I(p, n) <+ idt(g(V(p, n));
```

does not use any fewer state variables than the *ddt*-based implementation.

Before concluding this section, we briefly discuss the topic of current probes. As we have seen, probing voltage in Verilog-A is simple and straightforward. Probing current – although syntactically just as simple – requires a bit more care. Using the expression $I(n1, n2)$ on the right-hand side of a contribution statement yields the current flowing in the unnamed branch between nodes $n1$ and $n2$. Similarly, to probe the current through a named branch $br1$, the syntax $I(br1)$ would be utilized. If the branch being probed is not contributed to, the two terminals of the branch are effectively shorted together. Most Verilog-A

implementations will insert an extra state variable to probe the current through the branch, and thus it is desirable to avoid using current probes when possible. For example, the code fragment:

```
I(n1, n2) <+ V(n1, n2)/R;
x = f(I(n1, n2));
```

would typically be less efficient than the analogous:

```
I(n1, n2) <+ V(n1, n2)/R;
x = f(V(n1, n2)/R));
```

Current flow into module ports (terminals) can be probed with the expression *I*(<port_name>), where *port_name* is the name of the port. Note that it is usually an error to write *I*(port_name) instead, as the use of this expression on the right-hand side will create an unnamed shorted branch from *port_name* to ground, and thus almost certainly cause the model to behave in an undesirable way.

2.3. Analog Operators

As a general-purpose analog modeling language, Verilog-A includes a large number of "analog operators" that can be applied to signal waveforms. In addition to conventional operations such as differentiation, integration, and delay, the language also provides the transition, slew, circular integration, laplace transform, and Z-transform operators (see table below).

For compact model development, it is seldom necessary to use analog operators other than ddt. Occasionally, short delays may be used for some device applications, and laplace operators can sometimes prove useful for modeling passive circuitry or packaging outside the device. It is generally best to avoid the use of the integrator with initial conditions, the slew and transition filters, and the Z-transform operator because usage of these facilities restricts the range of analysis types that the model is suitable for [10]. Fortunately, there is almost never a need for such operations in compact modeling work.

Analog operators/ Waveform filters	ddt(x [,abs_tol])	Differentiate 'x' with respect to time.
	idt(x, [ic [, assert [, abs_tol]]])	Integrate 'x' with respect to time with initial condition 'ic.'
	idtmod(x, [ic [, modulus [, ffset]]])	Circular integration of 'x' with respect to time with initial condition 'ic' using modulus and offset.

transition(x [, delay [, rise_time [, fall_time]]])	Control details of signal transition expression 'x.'
slew(x [, max_pos [, max_neg]])	Control slew rate behavior of expression 'x.'
absdelay(x, time_delay, max_delay)	Output(t) = x(time − time_delay).
zi_nd(x, num, denom, period, [transition_time [,sample offset time])	z-domain filter function using numerator-denominator form.
zi_zd(x, zeros, denom, period, [transition_time [,sample offset time])	z-domain filter function using zero-denominator form.
zi_np(x, num, poles, period, [transition_time [,sample offset time])	z-domain filter function using numerator-pole form.
zi_zp(x, zeros, poles, period, [transition_time [,sample offset time])	z-domain filter function using zero-pole form.
laplace_nd(x, num, denom, [, abs_tol])	s-domain filter function using numerator-denominator form
laplace_zd(x, zeros, denom, [, abs_tol])	s-domain filter function using zero-denominator form
laplace_np(x, num, poles, [, abs_tol])	s-domain filter function using numerator-pole form
laplace_zp(x, zeros, poles, [, abs_tol])	s-domain filter function using zero-pole form

2.4. Noise

Noise analysis is an area of key importance for many analog applications, and thus comprehensive noise support is a requirement for compact model development. To this end, Verilog-A provides the *white_noise*,

flicker_noise, and *noise_table* functions.[1] For example, a noisy resistor would be modeled as:

```
I(p, n) <+ V(p, n)/R + white_noise (4*`P_K*$temperature*R);
```

whereas shot noise could be added via the expression:

```
I(b, c) <+ white_noise (2*`P_Q*Ic);
```

As we see above, the noise power arguments can be a function bias. A list of the available noise sources is given in the table below.

Noise functions	white_noise(power [, label])	Generate white noise of power 'power.' Contributions with the same label 'label' are combined for a module by the simulator.
	flicker_noise(power, exp [, label])	Generate pink noise of power 'power' at 1 Hz that varies in proportion to 1/f^exp. Contributions with the same label 'label' are combined for a module by the simulator.
	noise_table(vector [, label])	Generate noise where power is described by linear interpolation from vector 'vector' of frequency-power pairs. Contributions with the same label 'label' are combined for a module by the simulator.

In Verilog-A, each noise source is, by definition, independent. Correlation effects between noise sources can be modeled through linear combinations of real variables which are functions of the independent sources. For example, suppose that we would like to have a source $n1$ with power $P1$ and source $n2$

[1]The noise_table function may not be supported for some types of RF noise analysis. Its use should be avoided if possible.

with power $P2$, correlated to each other with a coefficient of K. This can be achieved by introducing three independent noise sources:

```
A = white_noise(K);
B = white_noise(P1-K);
C = white_noise(P2-K);
```

and then linearly combining them into the desired noise sources $n1$ and $n2$ as:

```
n1 = A+B;
n2 = A+C;
I(a, b) <+ n1;
I(c, d) <+ n2;
```

2.5. Analog Functions

Analog functions – sometimes referred to as user-defined functions – are directly analogous to their counterparts in conventional programming languages. Their primary role is to improve the readability and structure of a given analog block by encapsulating potentially complicated mathematical functionality. In some cases, using analog functions (instead of macros, for instance) can also lead to a smaller memory footprint.

Analog functions take as input a sequence of real or integer arguments, and return a real or integer value. In the 2.1 version of the standard, the arguments and return value are restricted to be scalar, and the arguments are passed by value. The 2.2 language standard allows the arguments to be arrays, and also allows them to be passed by reference (i.e., the function can effectively return values to the caller).

As an example of a typical analog function definition, we consider the following excerpt from a bipolar transistor model.

```
analog function real I_of_T;
   input IS, T, T_NOM, EG, N, Vth, XTI, XTB;
   real IS, T, T_NOM, EG, N, Vth, XTI, XTB;
   real ratioT;

   begin
     ratioT = T/T_NOM;
     I_of_T = IS / pow(ratioT, XTB) * exp((ratioT-1)*EG/
              (N * Vth))*pow(ratioT, XTI/N);
   end
endfunction // I_of_T
```

This function would be called from the module with the syntax

```
ISE_T = I_of_T(ISE, T, T_NOM, EG_T, NE, Vt, XTI, XTB);
ISC_T = I_of_T(ISC, T, T_NOM, EG_T, NC, Vt, XTI, XTB);
```

2.6. System Tasks

The Verilog-A language provides a set of "system tasks" which allow modules to interact with the simulation environment and with the input/output system. Each system task statement begins with the "$" character, followed by the name of the task and a parenthesized argument list. System tasks of interest to the compact model developer include the *$strobe* and *$debug*[2] calls for textual output to the display:

```
$strobe("V(coll) = %e", V(coll));
$debug("V(base)  = %e", V(base));
```

The *$strobe* task outputs its results at every converged solution point. In contrast, the *$debug* task generates output at every single Newton iteration, and (as its name suggests) is thus useful for debugging purposes. Numerous other system tasks are of course present in the language, and the reader is advised to consult [1] for a detailed list.

In addition to "system tasks", Verilog-A also provides several "system functions" (also occasionally referred to as system calls). These are distinct from system tasks in that they are function calls returning real-valued expressions, whereas the system tasks are statements that do not provide a return value. Of particular interest to compact model developers are *$temperature* (which returns the temperature in Kelvin), *$vt* (which returns the thermal voltage), and *$abstime* (which returns the current simulation time). Although compact models should not explicitly rely on time for their current-voltage characteristics, the *$abstime* call can be very useful for data display and debugging.

2.7. Conditional Statements, Looping Constructs, and Genvars

Conditional statements and for-loops are very useful language constructs for device modeling. Their usage in Verilog-A is similar to their usage in standard programming languages such as C, with one important distinction – analog operators (Section 2.3) may be used inside the body of these constructs only if the controlling expression is not a function of the state variables.[3] The restriction exists because analog operators must store their state internally, and thus need to monitor their arguments during the course of an entire analysis. To illustrate the restriction, consider the following simple code snippet:

```
if(V(ctrl) > 0) begin
   x = V(a, b);    // this is legal
```

[2]The $debug task is only present in the 2.2 (and later) versions of the standard.
[3]In the language of the standard, this is referred to as a "genvar expression".

```
    I(c, d) <+ V(a, b); // this is legal as well
    x = ddt(V(a, b)); // this is illegal:
                      // analog operator prohibited here
end
```

The entire code fragment above would be legal if V(ctrl) was replaced by a parameter type.

For-loops, although not as widely used as conditional statements, are still quite commonplace in compact modeling applications. They are particularly useful for such tasks as looping through the fingers of a multi-finger device or iterating through the emitters of a multiple-emitter transistor. For those situations where analog operators are needed within the body of a for-loop, the language introduces the "genvar" type of integer-valued variable. Variables which are declared as *genvar* may only be initialized within the controlling expressions of a *for*-loop statement, and may only be functions of static expressions (i.e., ones which are not functions of state variables, and are thus not dependent on bias). As an example of this concept, the code fragment below would insert a linear parallel RLC network into each of the *NF* fingers of a Verilog-A device:

```
electrical [1:`NF] no, ni;
real vk;
real [1:`NF] R, C, L;
// Code to initialize R/L/C arrays goes here...
genvar k;
for(k = 1; k <= `NF, k = k+1) begin
   vk = V(no[k], ni[k]);
   I(no[k], ni[k]) <+ vk/R[k] + C[k]*ddt(vk) + L[k]*idt(vk);
end
```

2.8. Hierarchical Module Instantiation

Verilog-A modules can be hierarchical in nature – each module may itself instantiate an arbitrary number of sub-modules. For example, if a detailed Verilog-A model for a bias-dependent junction capacitor has been written, a MOS model could instantiate instances of it with the statements

```
juncap #(.TRJ(TRJ1), .DTA(DTA1), ...  ) JUNCAPsource(BS, S);
juncap #(.TRJ(TRJ2), .DTA(DTA2), ...  ) JUNCAPdrain(BD, D);
```

The syntax above indicates that two juncap devices will be placed hierarchically within the parent module. The first of these would be named *JUNCAPsource*, and attached between nodes *BS* and *S*, while the second (named *JUNCAPdrain*) would be connected to nodes *BD* and *D*. Parameters are passed from the parent module to the child sub-device through a comma-separated list

after the '#' symbol. The values may themselves be functions of other parameters:

```
parameter real L = 0.1u from (0, inf];
parameter real W = 0.5u from (0, inf];
some_device #(.Area(L*W)) dev(n1, n2, n3);
```

2.9. Events and Memory States

As a general-purpose modeling language, Verilog-A includes some behavioral constructs that are best avoided in compact modeling applications. Chief among these are events (e.g., *@(cross)* and *@(timer)*) and the use of "memory states", which are further explained below.

The Verilog-A language standard mandates that local variables are initialized to zero at the beginning of the simulation, and that they retain their value after a given time point has converged. If a variable is used before it is assigned in a given module, it takes on the value from the previously-converged time point. We refer to such variables as "memory states". (In other literature [8], such variables may be referred to by other names, such as "hidden states".) Although such variables can be very useful for behavioral modeling applications, they clearly have very limited utility in compact model development. One possible use would be to limit the display of diagnostic information. For example, to print a warning only once, we could structure the code as follows:

```
integer warn_flag;
if(!warn_flag && R < 0) begin
   $strobe("Negative resistance in module %m");
   warn_flag = 1;
end
```

In addition to representing a questionable formulation from the standpoint of physical reality, modules with memory states can pose problems for RF simulation algorithms like harmonic balance and periodic shooting [9]. Indeed, for methods such as harmonic balance – which do not rely on conventional time-marching algorithms at all – the whole concept of memory states is particularly problematic. In the area of compact modeling most memory states can usually be attributed to an inadvertent mistake in variable usage, and compact model compilers should automatically warn the model developer of their presence [10].

3. Compact Model Development

3.1. Numerical Considerations

Circuit simulation algorithms are generally iterative in nature, and are typically based on the classical Newton-Raphson technique. To facilitate robust

convergence, semiconductor device models should have smooth, differentiable characteristics, and should guard against floating point exceptions that can occur during the course of iterating to a solution. It is important to remember that state variables can assume non-physical values during the course of the iterative process, and that "reasonable" nodal values are only guaranteed once the system has converged to a valid solution.

3.2. Model Topology

The program-flow aspect of the Verilog-A language tends to be straightforward, intuitive, and very similar to the languages that compact model developers are accustomed to. The contribution statements specifying model topology – while also fairly intuitive – do not have any direct counterparts in conventional programming languages, and consequently merit some additional discussion.

Most circuit simulators use Modified Nodal Analysis (MNA) [11] or something very similar to formulate the circuit equations. Consider a simple linear resistor, connected between nodes $n1$ and $n2$, represented by the constitutive equation $I = V/R$ (or $I(n1, n2) <+ V(n1, n2)/R$ in Verilog-A). A conventional circuit simulator would have state variables corresponding to nodes $n1$ and $n2$, and all components connected to these nodes would contribute terminal currents to the relevant Kirchoff's Current Law (KCL) equations. The resistor component would simply add the current V/R to one of the nodes, and subtract V/R from the other node. No additional equations or state variables would be necessary.

In contrast, consider a voltage source placed between nodes $n1$ and $n2$. The Verilog-A constitutive relation for this component takes on the form $V(n1, n2) <+ Vdc$, and in this case cannot be expressed in a "voltage-controlled" formulation. To handle this scenario, typical circuit simulators proceed to create a new variable representing the current through the source, and then add the equation $Vn1 - Vn2 - Vdc == 0$ as an extra row in the system.

The formulation method is important for compact model developers, since it can impact the size of the matrix. Because performance is a key issue in compact model development, it is important to have a good understanding of when "extra" state variables may be introduced by the various language constructs. The general rule of thumb (described in Section 2.2) is that contributing *to* a voltage (i.e., having $V(\ldots) <+$ on the left-hand side) or sensing a current (i.e., using $I(\ldots)$ within a right-hand side expression) can lead to the insertion of extra state variables.

One issue that frequently comes up in compact modeling work is the presence of optional parasitic resistors on the device terminals. These cannot be portably implemented using the standard contribution statement:

```
I(e, ei) <+ V(e, ei)/Re;
```

because the resistor value may be zero. The obvious solution – that is, using the voltage contribution:

```
V(e, ei) <+ Re*I(e, ei);
```

is portable and robust. However, as we have seen previously, this formulation will introduce an extra state variable for the resistive branch. For the case where the resistor value is zero, this results in the creation of not one but two extra state variables – one for the internal node *ei*, and one for the current through the voltage branch between nodes *e* and *ei*.

To overcome the aforementioned problem, modern compact model compilers will often make a special allowance for the following idiom:

```
if(Re > 0.0)
   I(e, ei) <+ V(e, ei)/Re;
else
   V(e, ei) <+ 0;
```

So long as *Re* is a static expression (i.e., one that does not depend on the values of the state variables) the model compiler will "collapse" the nodes *e* and *ei* into a single state variable if the parasitic resistance value is zero. In the case of implementations which do not special-case this construct, the code fragment will introduce a "switch branch" but will still execute correctly and be fully compliant with the language standard.

3.3. Compact Modeling Extensions

The recent 2.2 release of the language standard [1] has added several features of interest to compact model developers [7, 12]. Facilities have been added for explicit derivative access, portable output of local variables, efficient and convenient representation of parameter sets, more flexible specifications of user-defined analog functions, as well as several other common tasks. Table-based modeling support has also been added as a standard feature, enabling compact models to utilize table-driven characteristics.

Because the new standard has only recently been released, support for these features is not yet widely available across the various Verilog-A distributions. If portability is important, models utilizing the new feature set can check the predefined macro 'VAMS_COMPACT_MODELING; implementations that support the compact modeling extensions will have this definition present. Constructs that are dependent on the 2.2 feature set can be placed within an *'ifdef* for backward compatibility with earlier implementations.

A full detailed discussion of the new feature set is beyond the scope of this chapter; for more information, the reader is directed to the language standard [1]. Here, we present a brief overview of some of the more useful 2.2 functionality.

3.3.1. The ddx operator

Although Verilog-A compilers must internally compute symbolic derivatives to ensure that the Newton-Raphson process exhibits robust convergence, the language standard prior to version 2.2 did not allow model developers direct access to symbolic derivative information. This situation has been remedied with the introduction of the *ddx* operator. The *ddx* operator takes two arguments – the expression to be differentiated, and the state variable with respect to which the differentiation should take place:

```
Id = Area*Is*(limexp(V(pos, neg)/(n*$vt))-1);
Qd = tt*Id + Area*V(pos, neg)
     *Cjo/pow((1-V(pos, neg)/Phi), m);
Gd = ddx(Id, V(pos));
Cd = ddx(Qd, V(pos));
```

It is important to keep in mind that the derivative is a true partial derivative – that is, all state variables (i.e., nodal voltages and branch currents) except the one being differentiated with respect to will be held fixed. The state variables being held fixed should be distinguished from the local module variables, which may of course vary with the "with respect to" state variable.[4]

An important point is that differentiation with respect to a voltage difference is not allowed. For example, it may be tempting to calculate transconductance for a BJT as:

```
Gm = ddx (Ic, V(b, e)); // error!
```

This formulation has two inherent problems. The first of these is that differentiation with respect to a voltage difference is forbidden by the standard (as we saw above). The second issue is that nodes *b* (base) and *e* (emitter) in most bipolar models will represent the external device nodes, connecting to the intrinsic model only through the parasitic lead resistors. As such, partial derivatives with respect to these nodes will not yield the derivative derivative value that the model developer desires, since these unknowns are independent of the intrinsic device's nodal values under partial differentiation.

3.3.2. Output variables

Built-in devices in spice-like circuit simulators are typically able to output internal information such as small-signal operating point values for a given

[4]Note that this variation is only conceptual in nature; there is no numerical limiting process, since the standard specifies that the derivative should be "exact" in a symbolic sense.

model instance. To enable portable output of model-specific data, the 2.2 standard now mandates that module-scope variables with description and unit attributes should be output to the data set. For example, if the real variable *Gd* of the preceding section was declared as

```
(* desc = "diode conductance", units = "mhos" *) real Gd;
```

then the value of *Gd* would be made available for output and plotting at every solution point (including every time point of transient analysis).

3.3.3. Limiting functions

Circuit simulators have traditionally employed solution algorithms which effectively "limit" the potentially sharp changes in nodal values that can occur during the Newton-Raphson iterative process due to strong device nonlinearities. Prior to the 2.2 standard, the *limexp* operator was provided to fulfill this role when dealing with exponential junction nonlinearities. However, the *limexp* facility did not have the full generality available to built-in junction limiting algorithms, and was clearly not applicable to other forms of limiting that are sometimes used for various compact models.

To address this situation, the *$limit* facility was introduced into the 2.2 language standard. In addition to providing a flexible interface for user-specified limiting algorithms, the standard also recommends that simulation environments provide default implementations of the common "pnjlim" and "fetlim" algorithms, which presumably are used by the built-in (native) devices. This allows compact models written in Verilog-A to employ limiting algorithms that are consistent with their native counterparts. For more detailed information, the reader is referred to [7] and Section 10.9 of [1].

4. Examples

Verilog-A enables an efficient and fast process for compact model developers to create and distribute models. Already many popular models are available in Verilog-A format from a variety of sources (see table below). However, for this process to have wide acceptance, the experience of the end-user of the model must be much the same as it is with the current model distribution methods. That is, the model must be available in all the available analyses, the simulation results must be identical, and the simulation performance must not be impaired. This section illustrates how both industry-standard and complex models implemented in Verilog-A perform with an identical use-model as far as the end-user is concerned.

Model Type	Models
BJT	SPICE-GP, HiCUM, MEXTRAM, VBIC
MOSFET	BSIM3, BSIM4, BSIM5, BSIMSOI, MOS11, PSP, EKV, RPI-Shur TFT
GaAs FET	Angelov, Curtice, Parker-Skellern, TOM1/3

4.1. Angelov-Chalmers GaAs FET Model

The Angelov-Chalmers GaAs FET model is a prominent compact model used in high frequency circuit designs [13]. It delivers good representation of high power behavior while also providing good prediction of harmonics.

It is a relatively straight-forward model to code; however, it is not available in all simulators since it is used by only a small segment of the design community. The model requires less than four hundred lines of Verilog-A code, including parameter definitions; actual behavioral expressions are less than two hundred lines of code.

The I–V characteristics shown in Figure 1 compare the results of a simulation using the Verilog-A model to the simulator's built-in version. As can be seen, the results are identical, as one would expect. From the user's perspective the model behaves and performs as though it were a natively coded model. Besides the advantage of easy-access to the code for modifications and extensions, a Verilog-A implementation of the Angelov-Chalmers model provides access to the model in simulators where the vendors have not provided a native version.

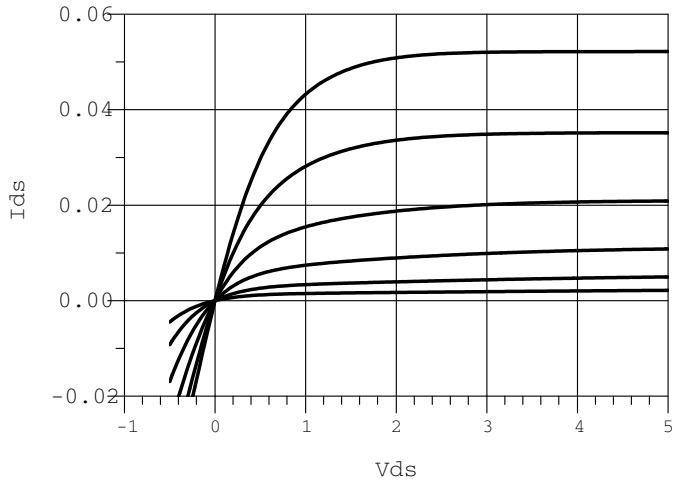


Figure 1. I–V characteristics of Built-in and Verilog-A versions of Angelov FET model.

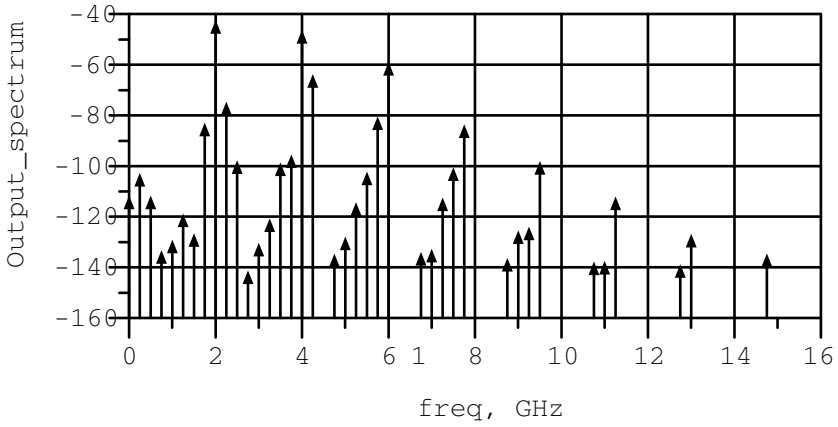


Figure 2. Output spectrum of EKV mixer analyzed in a commercial simulation program that does not natively support the EKV model.

4.2. EKV Model

The EKV model is another example of a popular MOSFET model that has been implemented in many, but not all, simulators. However, a Verilog-A version was released by the developers and this provides access to the model in simulators where it has not been implemented. Figure 2 illustrates a frequency domain simulation of an EKV mixer in a simulator that does not natively support the EKV model.

4.3. SPICE Gummel-Poon BJT

The SPICE Gummel-Poon BJT model is provided in virtually every analog simulator. Even though developed a half-century ago, until recently it has been general enough to sufficiently model advances in device technology. New compact models have been developed to address these improvements in topology and scaling. These recent models are more complicated, requiring more effort to extract the model parameters and using more simulation resources during analysis. However, in many cases minor modifications to the Gummel-Poon model would still be sufficient to accurately predict circuit performance. Verilog-A implementations of the BJT model allow users to add only the necessary behavior without adding unnecessary complications. For example, self-heating is an effect that is included in all of the next generation BJT models. It is important for devices used for power generation, or in materials with poor thermal characteristics. The self-heating effect can be modeled with just a few lines of Verilog-A code. A similar implementation in C-code, assuming the

end-user had access to the code, would be much more involved as it would be up to the developer to provide the numerous associated thermal derivatives for the simulator.

To demonstrate how Verilog-A models can be used in any analysis type, including frequency domain simulations such as harmonic balance, a real-world circuit using both Verilog-A compact models, Verilog-A behavioral models, and native simulator models for a modulator and demodulator.

Figure 3 shows the circuit schematic layout while the associated output is presented in Figure 4. The magnitude of the output is plotted along with the output for the Verilog-A model when self-heating is enabled.

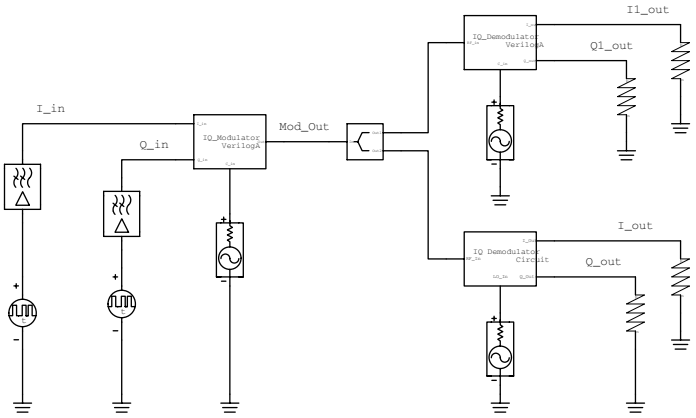


Figure 3. Schematic for a modulator-demodulator circuit employing Verilog-A for both compact models and behavioral models.

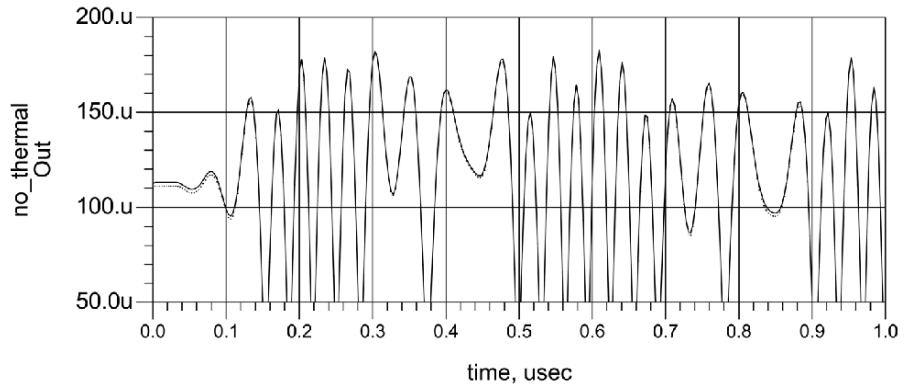


Figure 4. Output of demodulator in the time domain for a conventional SPICE Gummel-Poon model compared to the same model with a self-heating thermal circuit.

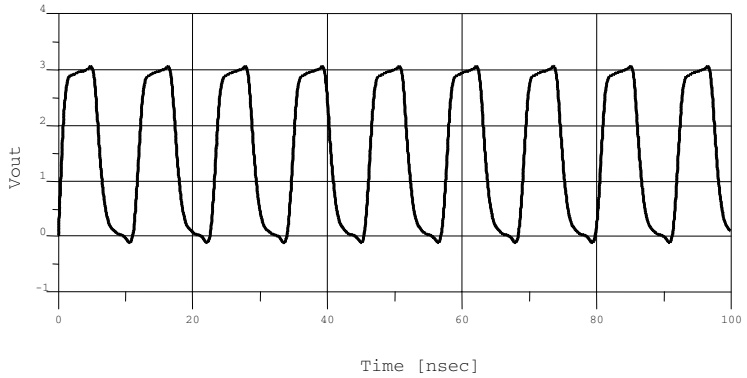


Figure 5. Ring oscillator output for both a native BSIM3 model and its Verilog-A equivalent.

4.4. BSIM3 MOSFET Model

The BSIM3 MOSFET model is the most extensively used compact model for analog and digital designs. It is the third generation model of the BSIM family and was developed with the intent of providing good fit to the underlying process as well as good mathematical behavior with respect to convergence. It is a complicated model with tens of thousands of lines of C-code and with hundreds of parameter values. In comparison, the Verilog-A implementation requires about one tenth the number of lines of code.

Since the model equations' derivatives are automatically generated, there is less chance of coding errors. This helps to accelerate the time it takes to get complex models out to the end-user. With shrinking geometries and novel device topologies, it is more difficult for any one compact model to accurately portray the device characteristics. Verilog-A allows new models to reach the end-user quicker; and for end-user feedback to return back to the model developer for model improvements.

Ring oscillator circuits are a simple way to exercise the model in a nonlinear manner. Small deviations in the models will result in large changes in the frequency of operation. Figure 5 shows the output of a ring oscillator for the built-in BSIM3 model and the Verilog-A equivalent model. As can be seen, the C-coded and Verilog-A models perform virtually identically.

References

[1] Verilog-AMS Language Reference Manual, Version 2.2, Accellera International, Inc.
[2] VHDL 1076.1 Language Reference Manual, IEEE.
[3] Lemaitre, L.; McAndrew, C.; Hamm, S. "ADMS – automatic device model synthesizer", *Proc. IEEE CICC*, **May 2002**, 27–30.

[4] Kundert, K. "Automatic model compilation – an idea whose time has come", *The Designer's Guide*, **May 2002**, (http://www.designers-guide.com).

[5] Mierzwinski, M.; O'Halloran, P.; Troyanovsky, B.; Dutton, R. "Changing the paradigm for compact model integration in circuit simulators using Verilog-A", **February 2003**, 376–379.

[6] Troyanovsky, B.; O'Halloran, P.; Mierzwinski, M. "Portable high – performance models using Verilog-A", *IEEE Conf. MTT*, **June 2003**.

[7] Coram, G.J.; "How to (and how NOT to) write a compact model in Verilog-A", *Proc. BMAS 2004*. CA: San Jose, **October 21-22, 2004**, 97–106.

[8] Kundert, K. "Hidden State in SpectreRF", *The Designer's Guide*, **May 2003**, (http://www.designers-guide.com).

[9] Kundert, K.; White, J.; Sangiovanni-Vincentelli, A. *Steady-State Methods for Simulating Analog and Microwave Circuits*. Kluwer Academic Publishers, **1990**.

[10] Troyanovsky, B.; O'Halloran, P.; Mierzwinski, M. "Analog RF model development with Verilog-A", *IEEE Radio Frequency IC Sympos.*, **June 2005**.

[11] Kundert, K. *The Designer's Guide to SPICE and Spectre*. Kluwer Academic Publishers, **1995**.

[12] Lemaitre, L.; Coram, G.; McAndrew, C.; Kundert, K. "Extensions to Verilog-A to support compact device modeling", *Proc. BMAS*, **October 2003**, 134–138.

[13] Angelov, I.; Zirath, H.; Rorsman, N. "A new empirical nonlinear model for HEMT and MESFET devices", *IEEE Trans. Microwave Theory and Tech.*, **December 1992**, *40(12)*.

INDEX