Xsim: A Compact Model for Bridging Technology Developers and Circuit Designers

Xing Zhou

School of Electrical & Electronic Engineering, Nanyang Technological University
Nanyang Avenue, Singapore 639798, exzhou@ntu.edu.sg

ABSTRACT

This paper describes the ideas and philosophy behind a new compact model (CM) for deep-submicron MOSFETs, called **Xsim**, which has been developed from scratch over the past few years. Similarities to and differences from existing popular models are pointed out. The opinions on many controversial debates in the CM field are given. The ultimate goal of the CM development in the context of technology/circuit modeling and optimization is outlined.

Keywords: Compact model, deep-submicron MOSFETs, parameter extraction, technology development, circuit simulation.

1 INTRODUCTION

The past few decades have witnessed a dramatic advancement in integrated-circuit (IC) technologies as a result of transistor scaling. Accompanied with this technological revolution is our improved understanding of the transistor physics and the ability to model it at various levels of abstraction. The popular circuit simulator SPICE as well as its associated generations of device models has played a key role in the development of the IC and electronic design automation (EDA) industries. Traditionally, technology developers and circuit designers are largely separate entities loosely linked by a set of GDS layout files and SPICE model parameters. This has been working fine before entering into the deep-submicron (DSM) era due to the fact that transistor characteristics are quite well modeled with unambiguous definition and extraction of its parameters, and statistical variations due to process fluctuations only represent a relatively small percentage of the nominal characteristics being modeled.

As the technology is scaled into the DSM regime, not only it becomes more difficult to model short-channel effects (SCEs) and higher-order physics in one unified compact form, but also increasingly important to relate the model characteristics to the inevitable process fluctuations such that the extracted model is statistically meaningful. There is also an increasing need to capture process effects on circuit performance and, inversely, optimize device performance by tuning the process. In this respect, physics-based compact models (CMs) are playing an increasingly important role in bridging the gap between technology developers and circuit designers.

To meet these new challenges, CM development is taking a few different directions. First, various efforts are being made to extend, modify, correct, or enhance the existing "regional" models that have been popularly used, such as use of "binning," introduction of new equations and effective quantities (e.g., various "electrical" gate oxide thickness), addition of large amount of fitting parameters, and mathematical tuning of parameters (e.g., global optimization, genetic algorithm). As a result, the model becomes more and more complex with unmanageable parameters for meaningful physical interpretation and prediction. A second approach that is becoming popular is the surface-potential-based model, which ends up with a single-region formulation with much less fitting parameters. While providing a physical solution to the surface potential of the intrinsic device, it is nontrivial to consider parasitic elements and SCEs as well as realistic doping profiles, besides its requirement for numerical iteration. Last but not the least, there are new methods in generating equivalent-circuit model from full numerical device simulations. No matter which approach, each one must have its own advantages over others as well as shortcomings. The final judgment will eventually depend on its scalability, predictability, and applicability in large circuit design and simulation.

It is phenomenal that, although higher-order effects (such as quantum-mechanical effect, velocity overshoot, or even discrete dopant fluctuation) should become important in DSM devices, the existing macro- or semi-empirically-based CMs still appear to work for the current technology nodes. Is this due to those higher-order effects being unimportant, or due to mathematical/empirical fitting used in the current models, or due to lack of alternatives? As pointed out by Tsividis [1] on analog circuit CAD: "What is needed is a model developed from scratch, by teams who truly understand device operation and appreciate the analog designer's needs at the same time. … Another difficulty is the entrenchment of established approaches and routines. … It would be a very difficult job to try and get it accepted by the community. It would have to be implemented correctly in several versions of Spice, appropriate parameter extraction systems would have to be developed, the foundries would have to be convinced to provide parameters for the new model, etc."

The question is: Do we need a new model developed from scratch? why? and how? It is with the belief that there is the increasing need to fill the gap between technology developers and circuit designers and refining existing models would make them more and more complex, that we

have started development of a new compact MOSFET DC model from scratch in the past few years, which we named it **Xsim**, with an associated parameter-extraction program called **DOUST**, which will be implemented in the multi-level circuit simulator, **XSIM**. This paper will try to answer why we need a new model and how we develop one. No attempt will be made to compare our model with other models since we believe that it is always easily "biased" when comparing one's own model (which would be used to its best potential) with others' (which may not even be properly used). Due to length limitations, this paper will focus on the ideas and approaches used in the course of **Xsim** development, with supporting data and plots to be presented at the conference and made available from the website [2]. Being an on-going development, the most recent "complete" model equations have been summarized in [3], with complete relevant references therein.

2 COMPACT-MODEL DEVELOPMENT

2.1 Mental Image of Reality

A model is a mental image of reality. One can have many different images of the same reality. A "dream" model for a MOSFET is to express the drain current (I_{ds}) as a function of geometry[1] (drawn gate length L and width W), terminal voltages (V_{gs}, V_{ds}, V_{bs}), temperature (T), and frequency (f): $I_{ds} = f(L, W, V_{gs}, V_{ds}, V_{bs}, T, f)$ such that the modeled I_{ds} as well as all of its higher-order partial derivatives can predict the measured one in the *full* range of the seven independent variables[2]. For long-channel MOSFETs, simple ideal MOS I_{ds} model can be as accurate as the most complicated models such as those from self-consistent Monte Carlo (MC) simulations. However, for DSM MOSFETs, it becomes extremely difficult to have one unified CM for accurate prediction in the full range of geometry and bias.

To model SCEs in DSM MOSFETs, new models are introduced and added to the existing ones, and with the addition of each SCE, new parameters are added. When the model still cannot cover the full geometry range, "binning" is used, which was described as a "clumsy and after-the-fact (but necessary) approach" [4] that effectively makes the model piecewise in geometry. The model parameter set effectively becomes n-fold as large for an n-binned model. In the limit of infinite number of bins, the model becomes a "single-device model" (i.e., nonscalable), which is only possible (and "necessary") for conventional single-device approach (i.e., model is fitted to bias at fixed geometry). This is analogous to "binning in bias": In the limit of large number of bias bins, a CM is equivalent to numerical device (or MC) simulation. (In this sense, a MC model can be considered as a complicated "compact" model; the same way as current CMs being viewed as "complex MC" by the logic-circuit designers.)

Current CMs consist of hundreds of fitting parameters to be extracted from a large set of measured I–V data. The reason of having to have so many parameters would be the model being either nonphysical or too complicated. This has been the motivation for us to develop a MOSFET CM from "scratch" that has minimum number of fitting parameters and minimum measurement data requirement. We believe that a MOSFET, no matter how small it is, is still a MOSFET that should follow the basic MOSFET physics. If a model, which accounts for all kinds of SCEs, cannot model long-channel devices well, it is then too complicated or ambiguous, if not wrong. To have a simple model, one must have a simple mental image of reality. We have been following Einstein's words as the guide in our CM development: "Everything should be made as simple as possible, but not any simpler."

2.2 Philosophy for CM Development

The experimental challenge in DSM technologies is not in making *one* "short-channel" transistor, but in making *millions* of the same transistor across different dies, wafers, and lots. Likewise, the challenge in DSM modeling is not just in accurate modeling of *one* transistor (perfect match to measurement), but mainly in *predictive* capability for a given technology as well as its associated process fluctuations. A major difference in our CM compared to conventional single-device (or "binnable") models lies in "technology" as opposed to "transistor" characterization. The threshold voltage (V_t), which is the key and most sensitive parameter in an I_{ds} model, in our CM is characterized by fitting to technology $V_t - L$ data rather than extracting from electrical I–V data, which implies that our CM is non-binnable.

Our approach to formulating CMs is to add SCEs step by step to the well-known long-channel equations, which is based on the belief that SCEs demonstrate themselves as a *gradual* change as the gate length alone is decreased. This means that the short-channel model should always converge to the simple one in the long-channel limit. When a higher-order effect is being added or calibrated, the parameters associated with that effect must be able to be "de-embedded" from the lower-order models that should have already been characterized. This idea is related to the sequence of model development and parameter extraction, which will be detailed in the following sub-sections.

In CM formulation, it is inevitable to introduce fitting parameters. In a CM, everything is "electrical" (effective) based on its definition. An "effective" quantity (or electrical equivalent) is only meaningful with the way it is defined. However, some parameters are supposed to be physical, such as gate oxide thickness (t_{ox}), LDD junction depth (x_j) and doping (N_{sd}), etc., and others are process-dependent. We separate the process-dependent *fitting* parameters ("unknown") from the process-variable *physical*

---

[1] Gate oxide thickness (t_{ox}) in the vertical dimension is often used as an independent physical input parameter.

[2] For reliability models, there is another independent variable, age.

parameters ("known" or estimated). The former is extracted at the average values of the latter, and then fixed in subsequent application of the model with the latter varied for statistical analysis of process fluctuations. When the fitting/physical parameters are used over all geometries, they may show different values at different bias conditions due to imperfectness of the model. In this case, we will introduce "bias coefficients" (analogous to temperature coefficients) to fit the model at corner bias conditions. When the fitting/physical parameters are used over biases, they may end up with different optimum values at different geometries. Then, we will model their geometry dependency semi-empirically. When too many empirical parameters have to be introduced, that is the indication that the model becomes unphysical and, thus, more efforts must be devoted to come up with new functionalities with less fitting parameters.

2.3 Sequence of CM Development

Ideal long-channel MOSFETs are well described by the simple equations when operated in linear, saturation, and subthreshold regions, respectively:

$$I_{ds0} = \mu_0 C_{ox} (W/L)\left[(V_{gs} - V_t)V_{ds} - ½A_b V_{ds}^2\right] \quad (1)$$

$$I_{dsat} = v_{sat} W C_{ox} \left(V_{gs} - V_t - A_b V_{dsat}\right) \quad (2)$$

$$I_{diff} = \mu_0 (W/L) v_{th}^2 C_d e^{(V_{gs} - V_t - V_{off})/(n v_{th})} \left(1 - e^{-V_{ds}/v_{th}}\right) \quad (3)$$

in which the threshold voltage is well modeled by

$$V_t = V_{FB} + \phi_{s0} + \gamma\sqrt{\phi_{s0} - V_{bs}} \quad (4)$$

$$\gamma = \sqrt{2q\varepsilon_{si} N}\Big/C_{ox} \quad (5)$$

$$\phi_{s0} = 2\phi_F = 2(kT/q)\ln(N/n_i) \quad (6)$$

μ_0 is the vertical-field mobility, A_b the bulk-charge factor, V_{dsat} the saturation voltage, N the channel doping, C_{ox} the gate capacitance, C_d the depletion capacitance, n the ideality factor, and v_{th} the thermal voltage, with well-known expressions. The effect of source/drain series resistance (R_{sd}) can be ignored for long-channel devices, thus, all the parameters (V_t, μ_0, W, L) are well defined and can be unambiguously measured. In scaled DSM MOSFETs, however, all these parameters, including V_{gs} and V_{ds} due to non-negligible voltage drop across R_{sd}, become coupled and must be treated as "effective" quantities for the *intrinsic* MOSFET, which have complex dependencies on the device *terminal* bias and *drawn* length and width.

Since there is no unambiguous definitions of the short-channel V_t, effective mobility μ_{eff}, and channel length L_{eff} as well as the effect of R_{sd}, our CM has been developed in the following sequence: (i) length- and bias-dependent V_t together with L_{eff} → (ii) μ_{eff} from long-channel $I_{ds} - V_{gs}$ at low V_{ds} → (iii) A_b from long-channel $I_{ds} - V_{ds}$ at high V_{gs} → (iv) R_{sd} from short-channel $I_{ds} - V_{gs}$ at low V_{ds} → (v) effective Early voltage V_{Aeff} from short-channel $I_{ds} - V_{ds}$ at high V_{gs} → (vi) subthreshold offset voltage V_{off} from short-channel $I_{ds} - V_{gs}$ at high V_{ds}. The followed sequence is such that those that have not been characterized will not be present (or have negligible effect) in the formulation of the current step being modeled.

We have spent considerable efforts in developing a physics-based V_t model that covers full range of geometry and bias, having a similar form to the long-channel one:

$$V_t = V_{FB} + \phi_s + \gamma_{eff}\sqrt{\phi_{s0} - V_{bs}} \quad (7)$$

in which SCEs are embedded in the surface potential (ϕ_s), effective body factor (γ_{eff}), and effective doping (N_{eff}), and all these effective quantities as well as V_t approach the long-channel expressions in the long-channel limit.

Short-channel V_t is known to depend on L_{eff}, μ_{eff}, and R_{sd}. In fact, conventional approach to L_{eff} and R_{sd} extraction from $1/I_{ds} - L$ at various $(V_{gs} - V_t)$ is ambiguous since partition of channel resistance is definition dependent and R_{sd} is nonscalable in DSM devices. Use of bias-dependent L_{eff} may complicate I_{ds} modeling at a later stage. With separate and physical modeling of V_t and R_{sd}, these two effects can be de-coupled, which is based on the assumptions that the voltage drop across R_{sd} is small at the threshold condition and the R_{sd} effect on V_t is contained in the $V_t - L$ data. With the constant-current definition of V_t calibrated at the long-channel reference current based on maximum-g_m V_t definition, minimum measurement data are needed for V_t model calibration over all gate lengths at corner bias conditions. Short-channel V_t is extracted together with a simple L_{eff} model for the metallurgical channel length including all major SCEs and RSCEs using an effective doping N_{eff} with an assumed Gaussian pile-up charge for the halo process. L_{eff} and N_{eff} are optimized through LDD lateral diffusion (σ) and halo centroid (l_μ) for minimum RMS error in all fitted $V_t - L$ data. The full V_t model resembles and approaches the long-channel one in the long-channel limit. This approach has been validated by numerical devices with various halo structures.

With physical L_{eff} and N_{eff} and accurate V_t modeling, I_{ds} model calibration becomes simpler. We have developed physics-based μ_0 (3 parameters), A_b (2 parameters), R_{sd} (2 parameters), V_{Aeff} (1 parameter), and V_{off} (1 parameter), models to be calibrated at the conditions when its respective effect is most pronounced. The resultant model covers the full range of gate lengths and bias conditions and has been verified with the 0.25-μm technology data. The final model has a single expression joining linear/saturation and strong/weak inversion regions by smoothing functions:

$$I_{ds} = \frac{I_{deff}}{1 + \left(R_{sd} I_{deff}\right)/V_{deff}} \quad (8)$$

$$I_{deff} = \left(1 + \frac{V_{ds} - V_{deff}}{V_{Aeff}}\right) I_{ds0} \quad (9)$$

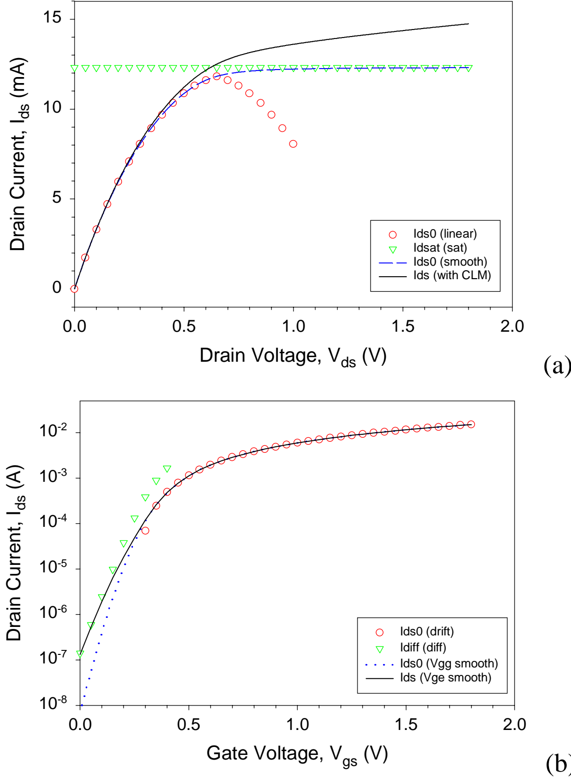

Figure 1: Single-piece model (8)–(10) (*solid lines*) smoothly joining the respective regional models (1)–(3) (*symbols*) in (a) $I_{ds} - V_{ds}$ from linear to saturation and (b) $I_{ds} - V_{gs}$ from subthreshold to strong-inversion regions.

$$I_{ds0} = \mu_{eff} C_{ox} \left(W / L_{eff} \right) V_{ge} \qquad (10)$$

where V_{ge} is a physically-derived gate/drain voltage product with smooth transition from subthreshold to saturation [5]. The above equations converge to Eqs. (1), (2), and (3) in the respective regions in the long-channel limit.

Fig. 1 illustrates our single-expression I_{ds} model as well as its asymptotes in different regions, which applies to all lengths and biases with one set of parameters. In our opinion, as long as the regional models are physically derived, use of smoothing functions is justified for the transition regions in which both drift and diffusion currents, or linear and saturation effects, are present. The major problem with the regional CM approach may be in the continuity of derivatives due to the smoothing functions. Recently, we have developed models for accurate g_{ds} prediction [6] including both velocity-overshoot and thermoelectric effects using energy-balance formulations, as well as improved modeling of A_b and μ_{eff} with accurate second (even third) order derivatives.

3 MODEL-PARAMETER EXTRACTION

`Xsim` is still under development. The preliminary version had 23 fitting parameters with a one-iteration extraction procedure. In recent enhancement, we have added parameters for long-channel drain-induced barrier lowering (DIBL), body-bias coefficients for the V_t model, V_{ds}-dependent bulk-charge factor, thermoelectric current in channel-length modulation (CLM) effect, as well as attempts to model the length-dependent saturation and subthreshold currents semi-empirically.

3.1 Measurement Data Requirement

While minimizing the number of fitting parameters, we also try to minimize the measurement data requirement. Although we have shown that the threshold voltage based on the "critical-current at linear threshold" definition [7] contains information on R_{sd} effect, we have adopted constant-current definition calibrated at long-channel maximum-g_m V_t, which can be readily available from automated electrical test protocols.

In principle, we need the following measured data for model-parameter extraction: long-channel $V_t - V_{bs}$ data at low and high V_{ds}; $V_t - L$ data at low and high V_{ds} for 2 (or 3) V_{bs} conditions [i.e., $4N$ (or $6N$) point (I, V) data where N is the number of different gate-length devices down to the V_t roll-off region]; long- and short-channel $I_{ds} - V_{gs}$ data at low V_{ds} and V_{bs} for μ_0 and R_{sd} calibration, respectively; long- and short-channel $I_{ds} - V_{ds}$ data at high V_{gs} and low V_{bs} for A_b and V_{Aeff} calibration, respectively; short-channel $I_{ds} - V_{gs}$ data at high V_{ds} and low V_{bs} for V_{off} calibration.

3.2 One- vs. Two-Iteration Extraction

In our previous CM, parameter extraction followed a one-iteration approach, which is based on the assumption that non-calibrated parameters have negligible effect in the current step of extraction and calibrated effect will not be affected by subsequent calibration. We also use simple equations before complete SCEs are being characterized. Errors are introduced in these assumptions due to switching of equations and "extreme" conditions being not large enough. This has been observed in our recent modeling of the 0.18-µm data (which was not obvious in the previous 0.25-µm data). In our new extraction approach, we have adopted a two-iteration scheme in which the first iteration parameter values are used as the initial guess for the second iteration, which uses the full short-channel model equations. Considerable improvement has been achieved.

3.3 Physical-Parameter Optimization

In our philosophy of separation of physical and fitting parameters, process-dependent fitting parameters (which also have their own physical meanings) are extracted at the average values of the process-variable parameters. These "fixed" physical parameters include t_{ox}, x_j, N_{sd}, CD correction (Δ_{CD}), and spacer thickness (S). However, certain physical parameters (such as LDD lateral diffusion σ, halo centroid l_μ, and saturation velocity v_{sat}) conceptually

exist but difficult to characterize. If they are used as fitting parameters in nonlinear regression, the CM model may not be easily converged, or may converge to unphysical values. In this case, these parameters are used in local optimization to obtain minimum RMS error in the target for which the respective parameters are defined. In our new extraction approach, σ and l_μ are used in $V_t - L$ optimization, and v_{sat} (together with the smoothing-function parameter) are used in $I_{ds} - V_{ds}$ and $g_{ds} - V_{ds}$ optimization.

3.4 Process Fluctuation and Prediction

Since our CM is physically calibrated to a given technology, we can apply the CM by "small" variations of the process-variable parameters around the extracted fitting-parameter values for monitoring process fluctuations in the current technology node; or by "large" variations of these parameters for prediction of the next-generation technology node. We have demonstrated the approach to correlating the CM parameter(s) to true process variable, such as implant dose and energy. This would ultimately provide a quick and reliable aid for technology developers.

4 MODEL IMPLEMENTATION AND APPLICATION

4.1 DOUST Parameter Extraction

The **`Xsim`** model has been implemented in an automated extraction program, **`DOUST`**, following the prioritized one- or two-iteration approach, with internal nonlinear regression from the model formulations. Measurement data (in popular formats) can be "converted," parameters "extracted," model "calibrated," and transistor characteristics "simulated" in a seamlessly way, as shown in the block diagram in Fig. 2.

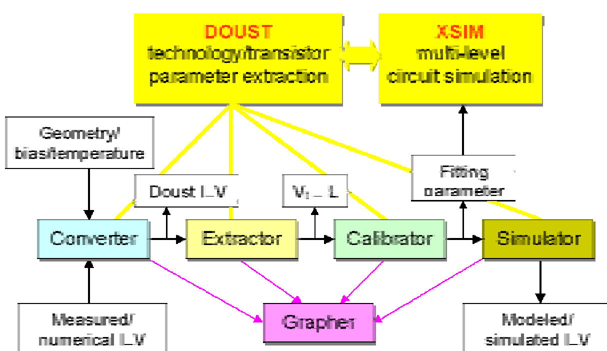

Figure 2: Block diagram of the **`DOUST`** parameter-extraction and **`XSIM`** circuit-simulation programs.

4.2 XSIM Multi-Level Circuit Simulation

The **`Xsim`** model as well as **`DOUST`** extraction will be linked to an implicit mixed-signal circuit simulator, **`XSIM`** [3], [8], which is based on subcircuit-expansion approach with automatic analog/digital circuit partition and dynamic mode switching. This allows system-level parameters to be linked to process variables in a consistent representation within a single-engine simulator. This development represents a first step towards bridging the technology developers and circuit designers, which would have significant impact on reducing design margins that translate into increased performance for a given technology.

5 SUMMARY AND CONCLUSIONS

Although many effects still have not been included in our CM (such as narrow width, substrate/gate currents, poly depletion, temperature effect, capacitance modeling, etc.), our approach to CM development has shown unique features and will provide a *guide* to technology developers, a *bridge* between the wafer fab and design house, and a *dynamic solution* to the future technologies and design methodologies. The ultimate goal is to have a CM for the technology developers, and a circuit simulator with true process variable input.

As there are so many versions of "*x*-SPICE," there exist quite a number of different "*y*-sim" models or simulators where "sim" normally implies "simulator." Whether the "**`X`**" in our acronym **`Xsim`** will be interpreted as "unknown" (as is usually used in math) or "excellent" (as an onomatopoeia) will ultimately depend on how well it applies to DSM technology/transistor/circuit modeling as well as how well it will be accepted by the community.

ACKNOWLEDGMENTS

Support of experimental data from Chartered Semiconductor Manufacturing Ltd. are gratefully acknowledged.

REFERENCES

[1] Y. P. Tsividis and K. Suyama, *IEEE J. Solid-State Circuits*, vol. 29, pp. 210–216, 1994.

[2] See complete information at the website: www.ntu.edu.sg/home/exzhou/research/DOUST/pub.htm

[3] X. Zhou, S. B. Chiah, K. Y. Lim, Y. Wang, X. Yu, S. Chwa, A. See, and L. Chan, *Proc. ICSICT-2001*, Shanghai, 2001, pp. 855–860.

[4] D. Foty, *MOSFET Modeling with SPICE, Principles and Practice*, Prentice Hall, 1997.

[5] K. Y. Lim and X. Zhou, "MOSFET subthreshold compact modeling with effective gate overdrive," *IEEE Trans. Electron Devices*, vol. 49, 2002.

[6] X. Zhou, S. B. Chiah, and K. Y. Lim, "A compact deep-submicron MOSFET g_{ds} model including hot-electron and thermoelectric effects," *Proc. ISDRS-01*, Washington DC, 2001.

[7] X. Zhou, K. Y. Lim, and D. Lim *IEEE Trans. Electron Devices*, vol. 46, pp. 807–809, 1999.

[8] S. S. Rofail, K.-S.Yeo, and K.-W. Chew, *IEE Proc. - Circuits, Devices and Systems*, vol. 146, pp. 83–89, 1999.