Unique Challenges in Compact Modeling

S. Mudanai, A. S. Roy
Logic Technology Development
TMG
Hillsboro, USA
Sivakumar.p.mudanai@intel.com

***Abstract*— Compact models are unique in the modeling domain in how accurate they need to be while maintaining significant computational efficiency. In this work, we will describe the requirements from a developer's perspective and illustrate the challenges in meeting these requirements and solutions using examples.**

Keywords—Compact Device Modeling, Body Charge model, Double Gate MOSFET, III-V model.

I. INTRODUCTION

Compact models form a critical link between the manufacturing teams and the chip design teams by mathematically capturing the properties of devices that are manufactured at the foundry for use in circuit simulators. These models capture every aspect of the device including AC, transient and noise in a framework that is fast enough for circuit simulations. These models also need to meet stringent requirements for successful use in industrial environments [1]. From a model developer's perspective these requirements, can be distilled as follows: (1) it should avoid regional approximations and provide continuous analytic expressions valid in all regions of operation, (2) it should be physics based and asymptotically correct and (3) it should support an intuitive physical framework that can be easily parameterized for accurate behavior over a range of operating conditions.

There are many challenges in meeting these requirements. They begin with the fundamental transport equations that are often in the form of Partial Differential Equations (PDEs) and hence cannot be programmatically simplified to formulate analytical and computationally efficient equations. The principal approach to overcome this challenge is to find the most suitable transformations and approximations so that the PDEs can be reduced to an Ordinary Differential Equation (ODE). The ODE can then be solved either analytically or numerically to formulate the core device equation using a core state variable. This approach is important in modeling the AC and transient behavior of the device. For traditional MOSFETs this approach has been refined over many years, where the long channel device equation serves as the core equation. The core state variable for these traditional MOSFET models tend to be either surface potential [2,3] or inversion charge [4,5]. For novel devices, however, the traditional approaches cannot be directly applied. In section II, we will showcase an approach to develop the surface potential equation for a double gate (DG) device with low effective mass.

After developing the core device equation, the next step is to incorporate all the additional physics needed to sufficiently describe the device, into the core model, as auxiliary models. The accuracy of a compact model is reliant on the ability of the auxiliary models to match the observed phenomenon. In addition these must be consistent with the core device equation and should also meet all of the three criteria. A common misconception is the assumption that common textbook like expressions, which are often derived for physical insight, can be easily incorporated into the core model. In Section III, we will illustrate these challenges in developing a body charge for a doped double gate device. In addition to accuracy a physically derived model can often provide insights, which cannot be gained by simply reading the results of TCAD simulation. We will illustrate this aspect in Section IV with the example of the development of a compact model for long change DIBL.

II. MODELING OF QUANTUM CAPACITANCE LIMITED DEVICES

The continued search for performance gains with technology scaling has led to the consideration of several alternatives to silicon such as GaAs, GaSb and InGaAs, which offer low effective mass for conduction. While the low effective mass results in high mobility, they also limit the charge available for conduction because of the low Density of States (LDOS). Considering that these materials will likely be used in thin body quantum wells for short channel control, we assume a double gate (DG) architecture for these devices, as shown in Fig. 1a.

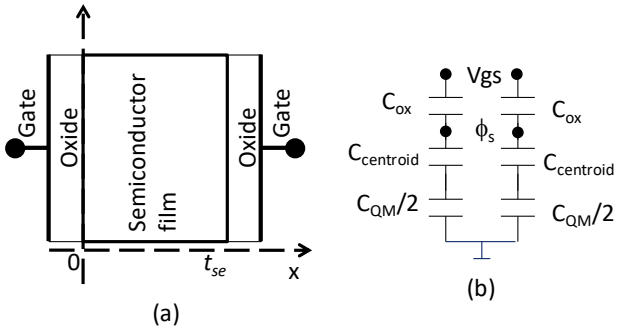

Fig 1: DG MOS structure and capacitance representation

From a modeling perspective the gate capacitance of a DG device can be modeled as shown in Fig. 1(b), where C_{QM}, is the quantum capacitance and $C_{centroid}$ is the charge centroid capacitance. In silicon the effective mass and the degeneracy

result in $C_{QM} >> C_{ox}$, hence modeling C_{ox} and C_{QM} is sufficient. Compact models for silicon DG devices solve the Poisson equation assuming Maxwell-Boltzmann distribution and account for $C_{centroid}$ by adjusting the effective oxide thickness. But, this approach does not work well for LDOS devices because they are degenerate and thickening the oxide will result in unphysical sensitivity to oxide thickness variations. Also, it is non-trivial to integrate the Poisson equation assuming Fermi-Dirac statistics in the semiconductor film. Hence, a two step approach [6] is used to solve for the surface potential. Firstly, the surface potential is solved by applying Gauss' law and equating the charge on the gate to the charge in the semiconductor film. Secondly, the charge distribution in the semiconductor film is used to estimate a centroid based correction term to the oxide thickness used in the first step.

The charge on the gate is derived using Gauss' law as follows,

$$Q_G = \varepsilon_{ox} E_{ox} = 2C_{ox}\left(V_g - V_{fb} - \phi_s\right), \quad (1)$$

where Q_G is the total charge on the gate, ε_{ox} is the oxide dielectric constant, V_{fb} is the flatband voltage, E_{ox} is the oxide field and ϕ_s is the surface potential. The charge in the semiconductor should be obtained by solving the Poisson equation. However, analytical solution of Poisson's equation assuming Fermi-Dirac statistics is non-trivial. Since, the semiconductor is thin, the potential will be assumed to be uniform across the semiconductor's thickness. The charge in an undoped semiconductor film is given by,

$$Q_s = \frac{q m_{De} KT}{\pi \hbar^2} ln(\xi_0 \xi_1) \quad (2)$$

$$\xi = 1 + exp\left[\frac{\phi_s - \frac{E_i}{q} - \frac{E_g}{2q}}{v_t}\right] - exp\left[\frac{-\frac{E_i}{q} - \frac{E_g}{2q}}{v_t}\right] \quad (3)$$

Where m_{De} is the electron density of states effective mass, k is the Boltzmann constant, T is the temperature, E_i is the energy of the ith electron subband, E_g is the bandgap of the material. For the sake of brevity only the electron charge contribution is shown in Eqn. (2). Equating the gate charge from Eqn. (1) to the semiconductor charge, we obtain the surface potential equation as,

$$2C_{ox}\left(V_g - V_{fb} - \phi_s\right) = \frac{q m_{De} KT}{\pi \hbar^2} ln(\xi). \quad (4)$$

The surface potential can be solved numerically using second order Newton iteration from equation (4). The solution for the gate capacitance obtained by solving Eqn. (4) is compared against numerical simulation data in Fig (2), where the compact model overestimates the gate capacitance. This is because Eqn. (4) is not accounting for the fact that the charge is distributed across the semiconductor. Instead it treats the semiconductor charge as an infinitely thin sheet charge at the oxide semiconductor interface. Hence, we need to consider the realistic distribution of the charge in the semiconductor.

The charge distribution in the first subband is nearly a sinusoid. Using this information we can obtain a relationship [6] between surface and center potential and the charge in the semiconductor, as

$$\phi_s = \phi_c + \frac{Q_s}{2C_{se}} \lambda \quad (5)$$

$$\lambda = \frac{1}{4} + \frac{1}{\pi^2} \quad (6)$$

where $C_{se} = \varepsilon_{se}/t_{se}$, is the semiconductor capacitance. Equating the charge on the gate to the charge in semiconductor and using (5),

$$2C_{ox}^{eff}\left(V_g - V_{fb} - \phi_s\right) = C_{QM} v_t \ln(\xi) \quad (7)$$

$$C_{ox}^{eff} = \frac{C_{ox}}{1 + \lambda C_{ox}/C_{se}} \quad (8)$$

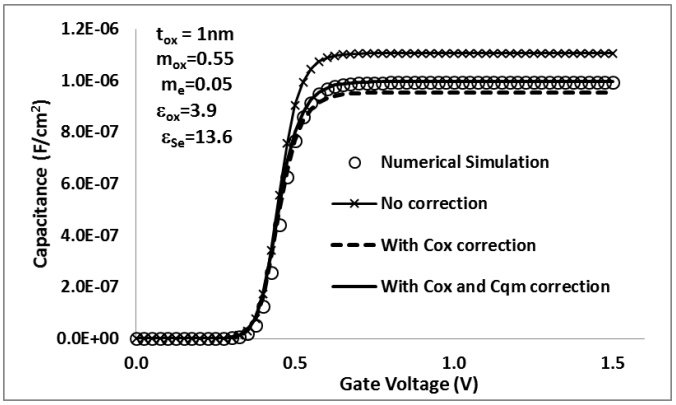

Fig 2: Gate capacitance for a 5nm thick LDOS material simulated numerically assuming no wavefunction penetration into the oxide compared to the model.

Comparison of the gate capacitance obtained using eqn. (7) against numerical simulation data, in Fig. 2, shows that the model now underpredicts the data. This is because quantum capacitance needs to be corrected for the perturbation of subband energy at high fields. Incorporating this perturbation [6] into eqn. (7), we obtain,

$$2C_{ox}^{eff}\left(V_g - V_{fb} - \phi_s\right) = C_{QM}^{eff} v_t \ln(\xi) \quad (9)$$

$$C_{QM}^{eff} = \frac{C_{QM}}{1 + \kappa C_{QM}/C_{se}} \quad (10)$$

where $\kappa = -1/24 + 1/8\pi^2$. The capacitance obtained by solving equation (9) matches numerical results very well.

Once we consider multiple subbands in the presence of wavefunction penetration, the approach presented so far become intractable. An intuitive semi-empirical expression is instead proposed that works not just for one subband but two subbands:

$$C_{ox}^{eff} = \frac{\varepsilon_{ox}}{t_{ox} + 0.7(t_{se}/4)\varepsilon_{ox}/\varepsilon_{se}} \quad (11)$$

The results of the gate capacitance using Eqn. 11, matches numerical simulation data for a wide range of body thickness and effective mass as shown in Fig. 3. The proposed expression can be understood as follows: When only one subband is occupied there are 3 effects to be considered: (1) the traditional charge centroid term (λ), (2) the correction for the quantum capacitance term (κ) and (3) the effect of wavefunction penetration. The latter two effects result in an increase in the increase of effective capacitance and are taken into account by the 0.7 factor in Eqn. (11). The t_{se}/4 factor accounts for the charge centroid correction. When the second subband is occupied, the quantum capacitance becomes larger than C_{ox}

and the correction based on charge centroid is sufficient. In compact modeling this type of approximation is critical in developing analytical functional forms which can then be used to develop the complete core I-V and C-V model [7].

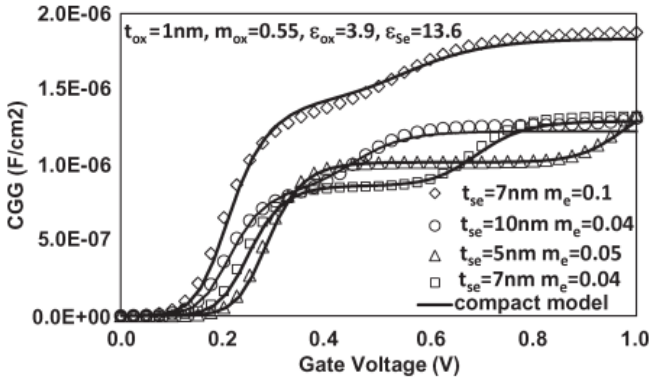

Figure 3: Compact model solution based on eqn. (11) compared against numerical solution for a range of semiconductor thickness and effective mass.

In summary, accurate compact modeling of low effective mass DG devices requires solution of the Poisson equation using Fermi statistics and the quantum effects which determine the charge distribution and the correction to quantum capacitance. The unique challenge is in using the insights, to develop a computationally efficient solution that is acceptable over the range of envisioned operation.

III. Body Charge Modeling in Double Gate Devices

Many models [2], [3], [8] have successfully used surface potential (SP) based modeling framework to sufficiently meet the compact modeling requirements in modeling planar MOSFETs. In the SP framework, the main challenges are (1) determining the body charge expression and linearization for drain current integration, and (2) determining the drain saturation voltage for solving the drain side surface potential. The original approximation for body charge modeling in a SP framework was proposed by Brews [9] as

$$Q_B = C_{ox}\gamma\sqrt{\phi_s - v_t}, \tag{12}$$

$$\gamma = \frac{\sqrt{2q\varepsilon_{Si}N_{dop}}}{C_{ox}}. \tag{13}$$

Where C_{ox}, is the oxide capacitance per unit area, N_{dop} is the body doping, ϕ_s is the surface potential and v_t is the thermal voltage. While eqn. (12) works well in in weak inversion, it is neither accurate nor amenable to continuous charge and capacitance model development through accumulation, depletion and inversion regions. Hence a modification to Q_B was proposed in [10]:

$$Q_B = C_{ox}\gamma\sqrt{\phi_s + v_t(e^{-\phi_s/v_t} - 1)}\ . \tag{14}$$

Equation (14) is valid in all regions of operation and provides asymptotically correct charge expressions for accumulation and depletion regions and serves as the backbone of most successful surface potential formulations.

The surface potential equation for a double gate MOSFET involves the solution of two non-linear system of equations for the surface (ϕ_s) and the center (ϕ_c) potential [11]:

$$V_{gs} - V_{fb} - \phi_s = \gamma\sqrt{F(\phi_s, V) - F(\phi_c, V)} \tag{15}$$

$$F(\phi, V) = \varphi + v_t e^{-\phi_s/v_t} + v_t e^{(\phi_s - 2\phi_B - V)/v_t}, \tag{16}$$

where V is the quasi-fermi level along the channel, V_{fb} is the flatband voltage and ϕ_B is the fermi potential due to body doping. Another equation is needed to constrain ϕ_s and ϕ_c, which is obtained from integrating the electric field from the surface to the center of the fin:

$$\phi_s - \phi_c = \frac{qN_{dop}t_{si}^2}{8\varepsilon_{si}}\left(1 + e^{(\phi_c - 2\phi_B - V)/v_t} - e^{-\phi_c/v_t}\right) \tag{17}$$

A more accurate relationship between ϕ_s and ϕ_c is derived in [11], which yields better accuracy. However, an analytic formulation of the body charge has not been reported in the literature thus far. Most models rely on treating the body charge in the full depletion limit, given by Eqn. (18), which works well in weak to strong inversion region of operation.

$$Q_B = \frac{qN_{dop}t_{Si}}{2}. \tag{18}$$

However, they need some patching functions to transition from full depletion to partial depletion to accumulation, which can make them vulnerable. Also any expression that is developed for body charge needs to be fully consistent with the core state variable formulation given by equations (15) and (17). If Q_B is not consistent with the core equations then the inversion charge which is obtained by evaluating the difference between in the gate charge and the body charge becomes inconsistent with the surface potential formulation. Unlike planar MOSFETs a charge sheet model inspired Q_B expression is not possible because low doped double gate devices are volume inverted in weak inversion.

In this work we propose the following body charge functional form:

$$Q_B = C_{ox}\gamma\sqrt{\frac{\phi_s - \phi_c + v_t\left(e^{-\frac{\phi_s}{v_t}} - e^{-\frac{\phi_c}{v_t}}\right)}{1 + e^{(\phi_c - 2\phi_B - V)/v_t}}}. \tag{19}$$

The expression (18) produces the correct asymptotic limits in partial depletion case where $\phi_c \rightarrow 0$, eqn (7) → eqn(3). In strong inversion where both ϕ_s and ϕ_c are much greater than v_t, one show that Eqn. (19) reduces to Eqn. (18).

In developing an expression for Q_B that is consistent with the surface potential equations (15), (16) and (17), we ensure a self-consistent model for body charge and inversion charge that is continuous and valid in all regions of transistor operation.

IV. Compact Modeling of Long Channel DIBL

In short channel devices DIBL is due to reduction of the barrier on the source side by the applied drain bias and the use halo implants reduces the field penetration and improves DIBL. In long channel devices, the use of halo implants actually increases DIBL and reduces the output impedance. This counter-intuitive phenomenon is critical to analog design which rely on high output impedance for both current mirror matching and amplifier gain. There is, however, limited literature available on this topic [12, 13].

The understanding begins with the three transistor equivalent circuit approach. In this approach the halo region on the source side, the low doped region in the middle and the halo region on the drain side are modeled as separate transistors in series. Though simulations using this approach reproduce all the unique features of halo implanted transistors, solving them analytically so that the essential physics and their limits can be understood and implemented in a compact model presents significant challenges.

In [12] each transistor is modeled as a Transformed Conductance (TC) G_v,

$$G_v = K_0 \frac{1}{L} exp\left(\widetilde{\Psi_s^{min}}\right). \quad (20)$$

Where K_0 has a weak dependence on the surface potential and can be considered as a constant, and $\widetilde{\Psi_s^{min}}$ represents the potential minimum offset by $3/2\phi_B$ and normalized by ν_t. Assuming a potential profile as shown in Fig. 4, the equivalent TC for the entire LC transistor can be derived as,

$$G_{eq} = K_0 \frac{exp(\widetilde{\Psi_h})exp(\widetilde{\Psi_m})}{L_m exp(\widetilde{\Psi_h}) + L_h exp(\widetilde{\Psi_m})(1+exp(-\Delta\Psi))}. \quad (21)$$

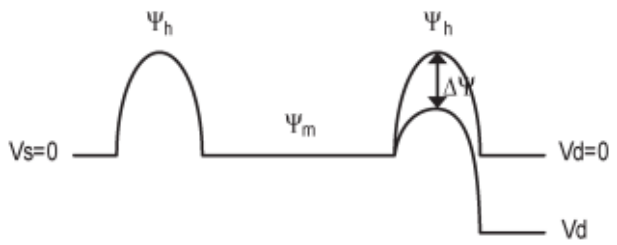

Figure 4: Equivalent energy profile of a long channel transistor. Ψ_h and Ψ_m represent the potential minimum for the halo and middle regions respectively. $\Delta\Psi$ is the effect of the drain bias on the halo transistor on the drain side.

Applying traditional MOSFET theory the shift in threshold voltage due to applied drain bias on the entire transistor [12] can be shown to be

$$\delta V_T = A\left(\log(2) - log\left(1 + e^{(-(BV_D + C\sqrt{V_D})/\nu_t)}\right)\right) \quad (22)$$

Where A is $n_{eq}\nu_t$ with n_{eq} representing effective subthreshold slope factor, while B and C are technology related parameters. Simulation result using Eqn. (22) matches experimental data very well as shown in Fig. (5) and Fig. (6) in [12].

To gain physical insight, consider the limit where the barrier reduction $\Delta\Psi$ to be >> ν_t and eqn (20) reduces to

$$\delta V_T = n_{eq}\nu_t \log(2) \quad (23).$$

This limit comes because, in the subthreshold region, the transport is determined by the maximum barrier height. When VD is zero, the current has to overcome two barriers. If each barrier represents a TC of *G*, the equivalent TC of the system is *G*/2. The application of a nonzero voltage at the drain terminal causes the drain barrier to decrease. As the TC has an exponential dependence on the surface potential, it becomes very high when the barrier gets lowered by a few ν_t and the total TC then becomes *G*. In other words, because of the barrier lowering at the drain, the equivalent TC is increased by factor of two, which when modeled as V_T results in Eqn. (23).

Eqn (22) and (23) provide critical insight into the mechanism by which halo implants degrade output impedance and provide an upper bound to the observable DIBL degradation that can be measured. The understanding of phenomenon which have a physical upper bound is specifically useful in the development of corner models for circuit design models. While an empirical auxiliary LC DIBL model would serve the purpose, the precise bias dependence on the drain bias can be a challenge without a physical derivation.

V. Conclusion

The challenges in compact modeling are wide and varied. In Section II, we showed that while analytical solutions exist for some sections of the problem space, it is non-trivial to develop an exact analytical solution for the full problem. However, it is possible to use the insights gained to develop analytical approximations that are sufficient. In section III we showed that even if trivial textbook expressions exist, finding an expression that is applicable across all regions of operation can still be challenging. These expressions must also be carefully constructed to be fully valid with the core model for a robust compact model. Similar challenges exist in the formulation of nearly every auxiliary model. In Section IV we show the importance of physically deriving the functional form for long channel DIBL and the resulting physical insight gained by using this approach as opposed to empirically modeling the effect. The uniqueness of the challenge in compact modeling is the need to comprehensively understand nearly every aspect of the device behavior, be able to model it a computationally efficient manner while meeting all the mathematical requirements for robust circuit simulation.

References

[1] R. Woltjer et. al., pp.1572-1580, vol. 51, Nov-Dec. 2007.
[2] G. Gildenblat et. al, pp. 1979-1993, vol. 53, Sept. 2006.
[3] M. Miura-Mattausch et. al, pp. 1994-2007, vol. 53, Sept. 2007
[4] J. -M. Sallese et. al., pp. 485-489, vol. 49, March 2005.
[5] J.-P. Duarte et. al., pp. 196-201, 41st ESSCIRC 2015.
[6] S. Mudanai et. al., pp. 4204-4211, IEEE TED, vol. 58, no. 12, 2011.
[7] A. Roy et. al., pp. 308-313, IEEE TED, vol. 61, no. 2, 2014.
[8] R. Rios et. al., pp. 755-789, IEDM 2004.
[9] J.R. Brews, pp. 345-355, Solid State Electron, vol. 21, 1978.
[10] J. Watts et. al., pp. 3-12, Tech. Proc. Workshop on Compact Modeling (2005).
[11] F. Liu et. al., pp. 3494-3502, IEEE TED, vol. 55, no. 12, 2008.
[12] A. S. Roy et. al., pp. 979-984, IEEE TED, vol. 58, no. 4, 2011.
[13] K. Cao et. al., pp. 171-174, IEDM Tech. Dig. 1999.