Ballistic MOS Model (BMM) Considering Full 2D Quantum Effects

Zhiping Yu, Dawei Zhang, and Lilin Tian

Institute of Microelectronics, Tsinghua University
Beijing 100084, China, yuzhip@tsinghua.edu.cn

ABSTRACT

As the channel length of MOSFETs is shrunk to below 50 nm, 2D quantum mechanical (QM) effects becomes profound on both the carrier confinement in the transverse direction to the channel and the carrier transport, mainly ballistic and tunneling, along the channel. A compact MOS model which incorporates the physics-based correction of 2D QM effects on the surface inversion charge density and ballistic transport is developed. The model has been applied to various MOS devices with gate length ranging from 45 nm to 14 nm and excellent agreement with published measurement data is achieved. It is proposed for the first time that WKB theory can be explored to model the subband lowering in the confined dimension because of the open boundary on the other dimension. An empirical formula for 2D-QM-corrected threshold voltage is provided. Mobility dependence on gate bias is investigated and modeled.

Keywords: compact MOS model, 2D quantum effects, ballistic transport, mobility modeling

1 INTRODUCTION

With the gate length of MOSFETs further scaled down into the sub-50nm regime, two non-classical physical mechanisms are generally recognized as playing a critical role in determination of device characteristics: ballistic transport and QM effects.

So far the efforts in modeling MOS with ballistic transport focus mainly on the regular device structures, such as the uniformly doped substrate in [1] for bulk CMOS and undoped silicon thin film as the channel region for double gate MOS in [2]. Practical devices, however, have highly nonuniform channel doping in order to optimize the device performance. In the meanwhile, with lateral (channel) dimension not much bigger than the size of carrier confinement in the transverse dimension, full 2D, not just 1D, QM effects have considerable impact on such parameters as the threshold voltage. Previous research was either based on 1D QM correction [3] to existing model formulation, or relies on 2D/quasi-2D [4] numerical simulation, the results of which are difficult to explore in developing physical insight of compact modeling.

In this work, a compact circuit model for sub-50 nm bulk silicon MOSFETs applicable to realistic channel doping is developed. The two aforementioned physical mechanisms, namely, the ballistic transport and the full 2D QM effects have been incorporated in the model. To emphasize the role of ballistic transport, the model is named as the Ballistic MOS Model, or BMM. An ingenious method based on the quantum mechanical WKB theory is proposed to transform the open boundary condition and the detailed potential profile along the channel to the subband lowering in the carrier confinement dimension. Because of the 2D effects, it is concluded that the threshold voltage increase in a deep-scaled down MOSFET due to the QM effects is less severe than a strict 1D treatment predicts.

The number of parameters in BMM is small compared to the BRIM models and since many parameters are physics based, the extraction process for model parameters is straightforward.

2 MODELING APPROACH

The analytical formula of the ballistic theory is discussed first as the basis for the complete model. The surface charge density of the inversion layer at the peak of the potential barrier along the channel, $Q(0)$, is then modeled using the proposed 2D QM correction scheme.

2.1 Modeling of the Ballistic Theory

The expression for the drain current considering the ballistic transport is described in [2],

$$\frac{I_D}{W} = Q(0)\frac{1-r}{1+r}\left[v_T\frac{\mathcal{F}_{1/2}(\eta)}{\mathcal{F}_0(\eta)}\right]\frac{1-\frac{\mathcal{F}_{1/2}(\eta-U_D)}{\mathcal{F}_{1/2}(\eta)}}{1+\frac{1-r}{1+r}\frac{\mathcal{F}_0(\eta-U_D)}{\mathcal{F}_0(\eta)}} \quad (1)$$

where the back scattering coefficient

$$r = \frac{l}{l+\lambda},\ l = L\left(V_T\frac{\beta}{V_{DS}}\right)^{\alpha},\ \lambda = V_T\frac{2\mu}{v_T}\frac{\mathcal{F}_0^2(\eta)}{\mathcal{F}_{-1}(\eta)\mathcal{F}_{1/2}(\eta)}$$

and l is the scattering critical length, λ the mean free path, $V_T = k_BT/q$ the thermal voltage, L and W are the channel length and gate width, respectively, μ the mobility, v_T the thermal velocity. $\mathcal{F}_j(x)$ is the j-th order Fermi integral, V_{DS} the drain bias, $U_D = V_{DS}/V_T$, and

α and β are fitting parameters. $\eta = (E_{FS} - E_{\max})/k_B T$ where E_{FS} is the Fermi energy at the source contact and $E_{\max}$ is the energy for the bottom of the first subband at the channel location, $x_{\max}$, where the position-dependent (along the channel) subband bottom energy reaches the maximum. The value of η is implicitly determined by the gate bias, V_{GS} and we'll find it through $Q(0)$.

The key modeling issue in (1) is $Q(0)$, the surface inversion layer charge density at $x_{\max}$ and it is determined by the applied terminal bias. With 1D QM correction, it is obtained in [3] that

$$Q(0) = C_{ox} V_{g,\text{eff}} \tag{2}$$

where $C_{ox} = \epsilon_{ox}/t_{ox}$ is the gate-oxide capacitance per unit surface area and $V_{g,\text{eff}}$ is the effective gate voltage,

$$V_{g,\text{eff}} = \frac{2nV_T \ln\left[1 + \exp\left(\frac{V_{od}}{2nV_T}\right)\right]}{1 + 2nC_{ox}\sqrt{\frac{2\phi_S}{q\epsilon_{Si}N_{\text{sub}}}}\exp\left[\frac{V_{od} - 2(V_{GS} - V_{th} - V_{\text{off}})}{2nV_T}\right]} \tag{3}$$

where $V_{od} = V_{GS} - V_{FB} - \phi_S - Q_{\text{dep}}/C_{ox}$ In the above, n is the subthreshold swing parameter, V_{off} is a parameter related to the threshold voltage change in different operation regions [3]. ϕ_S is the surface potential when the gate bias equals the threshold voltage, Q_{dep} and V_{th} are the depletion region charge and threshold voltage, respectively. For details, see [3].

On the other hand, according to Natori [1], η is linked to $Q(0)$ through the following relation:

$$Q(0) = \frac{qkT}{2\pi\hbar^2}\sum_{\text{valley}}\sum_i \sqrt{m_x m_y}\left[\ln\left(1 + e^{\eta}\right) + \ln\left(1 + e^{\eta - U_D}\right)\right]$$

$$\approx 2.5\frac{qkT}{2\pi\hbar^2}\sqrt{m_t m_l}\left[\ln\left(1 + e^{\eta}\right) + \ln\left(1 + e^{\eta - U_D}\right)\right] \tag{4}$$

where "valley" refers to counting all degenerate valleys, i counting all subbands, m_x effective mass in the transport direction and m_y in the direction of the device's width. For the substrate on (100) plane and consider only the lowest subband, an approximation can be obtained as in (4) [1], where m_t and m_l are transverse and longitudinal effective masses, respectively. Thus, knowing $Q(0)$, η can be solved for and the drain current can be evaluated in (1) from V_{GS} and V_{DS}.

2.2 Modeling of 2D QM Effects

It is long realized that the open boundary condition along the channel will affect the 1D solution to the Schrödinger equation in the transverse direction. Since the threshold voltage is largely determined by the ground energy level in the surface quantum well in this transverse direction, the proper assessment of the effect of potential distribution along the channel is crucial to the accurate calculation of the threshold voltage considering the 2D QM effects.

Even though the numerical 2D QM simulation can provide the information of the surface charge density at the x_{max}, which is close to the source end of the channel, it can hardly be used to develop a compact model. We propose a simple correction scheme to the subband edge obtained from the 1D Schrödinger equation using the solution of 1D Schrödinger equation along the channel. The solution in the channel direction is provided by the WKB theory for an arbitrary potential profile. Following Natori's coordinate system [1] of $x-$axis along the channel, $y-$axis on channel width, $z-$axis into the depth of the substrate, one can write the solution to the 3D Schrödinger equation using the separation of variables as follows.

$$\Psi(x,y,z) = \underbrace{\frac{A}{\sqrt{p(x)}}\exp\left(\frac{i}{\hbar}\int^x p(\xi)d\xi\right)}_{X(x)} \cdot \underbrace{\sqrt{\frac{2}{W}}\sin\left(\frac{n_y\pi}{W}y\right)}_{Y(y)} \cdot \underbrace{\varphi_{n_z}(z)}_{Z(z)} \tag{5}$$

In deriving the above wavefunction, we have assumed that the along the width of the channel, there is a flat quantum well with infinitely high barrier at $y = 0$ and $y = W$. $\varphi_{n_z}(z)$ is to be solved numerically as is done in 1D Schrödinger/Poisson equation solver. The key for the above solution in terms of our modeling need is the wavefunction along the x-direction. With the WKB theory,

$$p(x) = \sqrt{2m_x[E - U(x,y,z)]} \tag{6}$$

which may be an imaginary number if $E < U$. If U is a constant, then the solution is reduced to a plane wave and the energy E can never be below U and the kinetic energy $\hbar^2 k^2/2m_x$ is always positive. However, the potential distribution along the channel is not uniform and peaks at $x_{\max}$.

We define the subband edge as the energy when the wavefunction in the x-direction becomes localized, which means that we extend the solution range to include the tunneling regime as well. Since when $E < U$, the quantity of $(i/\hbar)\int^x p(\xi)d\xi$ becomes negative, indicating the wave is decaying within the potential barrier. Thus one can define the "locality" as that the exponential term in $X(x)$ becomes negligibly small while integrating from the left classical turning point ($x = a$) to the right ($x = b$) of the barrier. The energy level which demarcates the boundary between the nonlocal and local states can be found by solving the following nonlinear equation numerically.

$$\frac{i}{\hbar}\int_a^b \sqrt{2m_x[E_d - U(x)]}dx = -4 \tag{7}$$

for $e^{-4} = 1.83\%$ can be considered as negligibly small. From the solved E_d one can further define $\Delta V_{2\mathrm{D-QM}} = (U_0 - E_d)/q$ where U_0 is the peak value of the potential $U(x)$. This ΔV will be used as the 2D QM correction term in the 1D QM formulation for the threshold voltage in the existing model.

2.3 Model for Threshold Voltage

The V_{th} in (3) will be modified using the corrections due to the full 2D QM effects, the short-channel-effects (SCEs) and the drain-induced-barrier-lowering (DIBL) simply as

$$V_{th,\mathrm{eff}} = V_{th} - \Delta V_{2\mathrm{D-QM}} - \Delta V_{\mathrm{SCE}} - \Delta V_{\mathrm{DIBL}} \qquad (8)$$

where ΔV_{SCE} and ΔV_{DIBL} are taken from BSIM 4. This update of V_{th} completes the modeling of $Q(0)$ and hence I_D.

2.4 Mobility Model

With varying channel length in the sub-50nm range and nonuniform channel doping, the accurate mobility model is critical to the overall model accuracy. The gate bias dependence of the mobility is investigated and modeled. The following formula is used for mobility degradation

$$\mu(V_{GS}) = \frac{\mu_0}{1 + \theta_1 V_{GS} + \theta_2 V_{GS}^2} \qquad (9)$$

where μ_0, θ_1, and θ_2 are fitting parameters.

3 PARAMETER EXTRACTION

Overall, there are seven model parameters which need to be determined to evaluate $I_D(V_{GS}, V_{DS})$. They are t_{ox} and N_{sub} for $V_{g,\mathrm{eff}}$ (hence for $Q(0)$ as well) in (3), α and β in (1), and μ_0, θ_1, and θ_2 in (9). To determine these parameters for a specific MOSFET, at least one transfer curve and three output curves from experimental data are needed. In the process of parameter extraction, first, t_{ox} and N_{sub} are adjusted using the effective oxide thickness (EOT) and the average channel doping as the initial values and setting the other five parameters to default values (α and β taken from [1], and parameters for mobility from MEDICI of Synopsys) to fit the transfer curve. As long as the error is smaller than 30%, the values of t_{ox} and N_{sub} are fixed temporarily. Second, to determine α and β, at least one output curve is needed. The values of α and β are sensitive to the shape of I-V curve and can be determined relatively easily from the measured data. Finally, three parameters for μ are extracted from three output curves with different gate biases. The above steps may need iterations to get the overall error within desired level. Our experience shows that error between analytical results and experimental data can be as low as 5%.

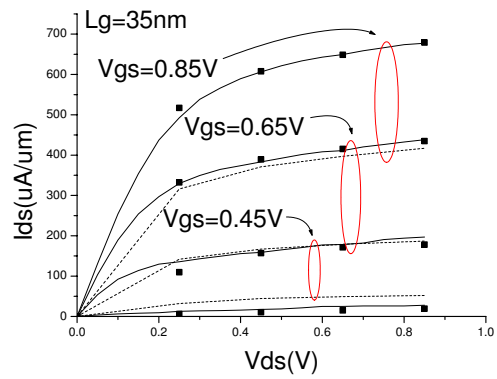

Figure 1: Comparison between simulation (lines) and measurement (symbols) data for Toshiba 35nm CMOS [6]. Dashed line is without 2D QM correction to V_{th}.

4 RESULTS

The compact model is used to calculate the current of the bulk silicon devices in [5]-[7], having respective gate lengths of 15nm, 35nm, and 42nm. The devices are all made either in industry or in the laboratory. The fitted parameters for these three devices are listed in Table 1. The values of N_{sub} are physically reasonable compared to the real doping in [5]-[7] and the values of t_{ox} are also almost the same as their EOTs. β is constant across the devices while α monotonically decreases with the increase of the gate length.

In Fig. 1, the comparison of the output characteristics of the 35nm device in [6] between experimental data and analytical results is presented. The analytical results without modeling ΔV_{th} are also compared to the experimental data. It is clear from Fig. 1 that without considering the QM effects along the channel, the analytical results grossly underestimate the real data, indicating that the strict 1D QM correction in the substrate exaggerates the increase of the threshold voltage due to the QM effects. In Fig. 2, the comparison of the transfer characteristics of 15nm device in [5] between experimental data and analytical results is presented. From Figs. 1-2, it can be seen that the analytical results agree well with the experimental data, proving that this compact model is applicable to those devices with a sub-50nm feature size. Our preliminary investigation even shows that the smaller the device size, the better the fitting accuracy of the BMM model.

5 CONCLUSIONS

A compact model for sub-50 nm MOSFETs has been developed. Advanced physics such as the ballistic transport and 2D quantum mechanical effects are incorporated in the model. The model is applicable to the devices with nonuniform channel doping. The model

Table 1: Parameters extracted for three bulk CMOS devices with gate length of 15, 35, and 42nm.

L_g (nm)	t_{ox} (nm)	N_{sub} (cm^{-3})	α	β	μ_0 (cm^2/V·s)	θ_1 (V^{-1})	θ_2 (V^{-2})
15	1	9×10^{18}	0.3	1.08	60.976	-6.240	11.266
35	1.2	2.3×10^{18}	0.2	1.08	167.504	-5.551	11.568
42	2	1×10^{18}	0.18	1.08	88.81	-3.352	5.711

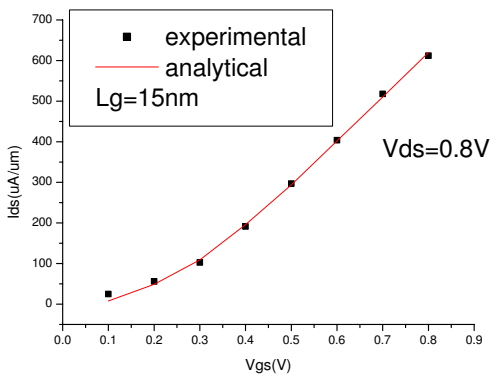

Figure 2: Transfer curve from modeling (solid line) and measurement (symbols) data for AMD 15nm CMOS [5].

has been tested using several published device structures with gate length ranging from 45 nm to 14 nm and the comparison between the simulated and measured data shows the model has good accuracy and predictability.

6 ACKNOWLEDGEMENT

This work is supported by a research grant (973) from Ministry of Science and Technology in China.

REFERENCES

[1] K. Natori, "Ballistic metal-oxide-semiconductor field effect transistor," *J. Appl. Phys.*, Vol. 76, 4879-4889, 1994. K. Natori, "Scaling limit of the MOS transistor – A ballistic MOSFET," *IEICE Trans. Electron.*, Vol. E84-C, 1029-1036, Aug. 2001.

[2] A. Rahman and S. Lundstrom, "A compact scattering model for the nanoscale double-gate MOSFET," *IEEE Trans. Electron Devices*, Vol. 49, 481-489, March 2002.

[3] Yutao Ma, Litian Liu, Lilin Tian, Zhiping Yu, and Zhijian Li, "Analytical charge control and I-V model for submicron and deep-submicron MOSFET's fully comprising quantum mechanical effects," *IEEE Trans. Computer-Aided-Design of Integrated Circuits and Systems*, Vol. 20, No. 4, 495-502, April 2001.

[4] A. Pirovano, A. L. Lacaita, and A. S. Spinelli, "Two-Dimensional quantum effects in nanoscale MOSFETs," *IEEE Trans. Electron Devices*, Vol. 49, 25-31, Jan. 2002.

[5] Bin Yu, Haihong Wang, Amol Joshi, Qi Xiang, Effiong Ibok, and Mingren Lin, "15nm gate length planar CMOS transistor," *IEDM Tech. Dig.,* 11.7.1-11.7.3, Dec. 2001.

[6] S. Inaba, *et al.*, "High performance 35nm gate length CMOS with NO oxynitride gate dielectric and Ni salicide," *IEDM Tech. Dig.,* 29.6.1-29.6.4, Dec. 2001.

[7] Qiuxia Xu, He Qian, Zhengsheng Han, Ming Liu, Ruibing Hou, Baoqing Chen, Haojie Jiang, Yuyin Zhao, and Dexin Wu, "High performance 42nm gate length CMOS device," *Chinese Journal of Semiconductors,* Vol. 24, Supplement, 153-160, April 2003.