Modeling of MOS Scaling with Emphasis on Gate Tunneling and Source/Drain Resistance

Chang-Hoon Choi, Zhiping Yu, and Robert W. Dutton
Center for Integrated Systems, Stanford University, Stanford, CA 94305

1 Introduction

With remarkable improvements in CMOS of gate length of 70 nm and operating voltage at less than 1 V, an oxide thickness of less than 1.5 nm is needed, which corresponds to 2–3 layers of silicon atoms. With such a thin oxide, the direct tunneling results in exponential increases in gate leakage current, increased power consumption and deteriorated device performance [1]. Another issue is the loss of inversion charge and transconductance due to the quantization of carriers in the channel as well as polysilicon gate depletion effects. In the QM model, the inversion charge profile peaks at around 10 Å below the silicon surface such that gate capacitance and inversion charge are both effectively reduced to those of an equivalent oxide a few angstroms to nanometer thicker than the physical oxide.

From the device technology point of view, shallow junctions in the Source/Drain extension (SDE) regions are needed to minimize the short channel effects in scaling down, sub-100 nm MOSFETs. Even though the SIA road-map predicts the SDE junction depths (X_j) as low as 10 nm to achieve an $L_{eff} = 50$ nm, too shallow an X_j can lead to a high external resistance, which in turn results in I_{dsat} degradation [2]. This implies that further improvements in I_{dsat} and MOSFET performance cannot be expected when the channel resistance becomes comparable to the source and drain resistance. In other words, further down-scaling without improvement of MOSFET performance is meaningless, even if short-channel effects are completely suppressed by introducing the shallow SDE.

Therefore, special characterization is needed to account for above effects in the choice of design parameters, especially for future sub-100 nm MOS technology. In this paper, modeling of CMOS scaling issues in terms of gate characteristics and S/D resistance is presented.

2 Gate Characteristics and Source/Drain Resistance

2.1 Gate Characteristics in Ultra-thin Gate Oxide

Fig. 1 shows measured C–V curves of the n^+−poly / SiO_2 / p−Si MOS capacitors for oxide thicknesses of 1.3, 1.5, and 1.8 nm, respectively. The capacitor area is 100 μm × 100 μm, and the small signal frequency is 100 kHz. It can be observed that there are sharp transitions in the capacitance both in the inversion and accumulation regions. Also, the peak capacitance for a gate oxide thickness, t_{ox}, of 1.3 nm is less than that of $t_{ox} = 1.8$ nm. This distorted capacitance behavior is caused by the gate tunneling current through the oxide. This anomalous C-V behavior is a result of interpreting the measured y−parameter in a conventional capacitance measurement system with the imaginary part of the admittance fully contributed by the gate MOS capacitance [3]. As a result, measured capacitance for oxides thinner than 2 nm shows qualitatively different behavior compared to thicker oxides. In order to model this behavior quantitatively, an accurate gate current calculation needs to be coupled with the QM effects.

2.2 Source/Drain Resistance in Shallow Source/Drain Junction

Even though the shallow junction of the Source/Drain Extension is a key element in scaling CMOS down to sub-100 nm, too high a series resistance can cause degradation of the driving current. Thus, the fundamental trade-offs in sub-100 nm CMOS scaling is between short channel effects and series resistance in SDE region. Accordingly, to minimize external resistance components it is necessary to predict them accurately as they relate to different process conditions as well as device structures. To determine the exact external resistance experimentally, the L_m-array, where L_m is the gate length, method is most commonly used. However, this method is based on a linear regression which possibly could extract wrong or even negative resistance values from the nonlinear R-L characteristics caused by process variations [4]. Moreover, the device designer cannot directly extract the different resistance components contributing to R_s and R_d from these methods since experimentally extracted resistance values are inherently lumped. For device optimization, understanding of the details among different resistance components with respect to the doping profiles, dimensions, and material parameters are of great importance.

3 Quantum Mechanical Modeling of C-V and Gate Tunneling Current

3.1 C-V Model with QM Corrections

For the modeling of gate characteristics an empirical and hybrid model for the QM correction in MOS structures is proposed and implemented in a 2D device simulator, which combines both the van Dort and Hansch models [5].

In the vicinity of the Si/SiO_2 interface, the density of conduction band states can be expressed as

$$N_C(z) = N_C[1 - e^{-(z+z_0)^2/\lambda_{th}^2}] \tag{1}$$

where the $z-$axis is perpendicular to the Si-SiO_2 interface and z_0 is introduced to account for the non-zero carrier concentration at the surface ($z = 0$) due to finite barrier height between Si and SiO_2. N_C is the normal density of conduction band states, which is position-independent, and λ_{th} is a thermal wavelength.

To eliminate the flatband singularity that occurs in the van Dort model, a novel bandgap broadening expression is used. The bandgap correction term, depending on the 2/3 power of the transverse surface electric field, F_S, in the van Dort model is given as

$$\Delta E_g(z) = \frac{\kappa\beta}{2k_BT} F_S^{2/3} \left[\frac{2e^{(z/L)^2}}{1 + e^{-2(z/L)^2}}\right] \tag{2}$$

where β is a physical constant, κ is the fitting parameter with the theoretical value of unity and L is a characteristic decay length. The singularity originates from the derivative of this quantity with respect to $F_S = 0$ in the gate capacitance computation. To eliminate this flatband singularity, the

$F_S^{2/3}$ term in Eq. (2) is now replaced by

$$\frac{F_S^2}{ae^{-(F_S/\sigma)^2} + F_S^{4/3}} \tag{3}$$

where both a and σ are adjustable parameters. This transformation preserves the asymptotic dependence of the bandgap correction on $F_S^{2/3}$ when $F_S \rightarrow \pm\infty$ while eliminating the flatband singularity.

Fig. 2 shows the comparison between experimentally measured and simulated nMOS C–V data for an oxide thickness of 2 nm and area of 100 μm $\times$ 100 μm. For this analysis, the substrate doping profile is obtained using a process simulation that includes TED (transient enhanced diffusion) effects [6]. The reduction of gate capacitance in the accumulation region is related to the QM effect; the reduction in the inversion region is related to both the QM and poly depletion effects.

3.2 Gate Tunneling Simulation

In this work the tunneling current is calculated using a 1D Green's function simulator, NEMO [7]. Because the multiple-scattering effects become extremely important, the WKB approximation is no longer satisfied for the thinner SiO_2 layers [8]. NEMO considers the injection from quasi-bound states and continuum states; hence an accurate calculation can be achievable for the density of states. Carrier density is calculated quantum mechanically in the device; scattering effects at the contacts can be approximated by the potential used in calculation of the surface Green's function.

Fig. 3 shows the measured and simulated gate current for three gate oxide thicknesses. The discrepancy between the measured and simulated data is probably caused by either surface roughness or uncertainty in determining the effective oxide thickness.

4 Modeling of MOS Capacitor with Ultra-thin Gate Oxide

4.1 Equivalent Circuit Modeling

To characterize the MOS C-V for gate oxide thicknesses ranging from 1.3–1.8 nm, an equivalent circuit is introduced. First, the silicon surface region is divided into small rectangular segments perpendicularly to the channel-current flow in order to consider the channel resistance, as in Fig. 4 (a). In addition, the polysilicon area is vertically divided into small pieces to consider the polysilicon resistance effect. R_{ch} represents the channel resistance for each segment; this channel resistance also varies as a function of gate bias. R_g represents a distributed resistance in the polysilicon gate, and the measured sheet resistance of the polysilicon is 5 $\Omega/\Box$, which is much smaller than the channel resistance ($R_{ch} \simeq 1.5$ K$\Omega/\Box$). Fig. 4 (b) is the equivalent circuit for the single vertical segment represented in Fig. 4 (a). Then the equivalent circuit for the whole device structure is constructed by incorporating all the vertical segments. The tunneling current is modeled using a nonlinear, voltage-controlled current source ($i = f(V_g)$) derived from the gate current simulation discussed previously. The current values of each segment are then scaled according to the number of RC stages used. Also, the gate capacitance without the gate tunneling effect is modeled using a nonlinear, voltage-controlled capacitance ($C = f(V_g)$). These voltage-controlled elements contain one-dimensional and piecewise linear voltage-dependent capacitance and voltage-dependent current tables which are built from C-V and gate tunneling simulations, respectively. SPICE *ac* small-signal analysis is then performed to find the input admittance ($Y_{in} = G + j\omega C$) of the circuit at a given frequency of 100 kHz. The C–V data including gate tunneling effects is finally obtained by taking the imaginary component of Y_{in}. Fig. 5 compares the modeling with the measurements for oxide thicknesses of 1.3, 1.5 and 1.8 nm, respectively, when the number of RC segments is 20. When it exceeds five, the number of RC segments makes no difference in terms of the final capacitance values. The extracted channel resistance values in the inversion and accumulation regions are 2.5 and 1.1 K$\Omega/\Box$, respectively.

4.2 Channel Length Dependence of Gate Current

Gate tunneling current with respect to the drain bias (i.e. $V_d > 0$ V) can be modeled using the equivalent circuit, as shown in Fig. 6. Here, the channel length dependence of I_{gate} can be estimated by considering the short channel effects, i.e., drain-induced barrier lowering, due to the distributed channel resistance [9]. As a result, the simulated gate tunneling current decreases as the drain bias increases since the potential difference between the gate and the channel surface becomes smaller, as shown in Fig. 7 (a). Fig. 7 (b) shows the simulated gate-channel dependence of I_{gate}; I_{gate} decreases in an inverse proportion to $L_g^{1.5}$, which is comparable to the experiments by Momose *et al.* [10], where the slopes were 1.5 and 1.8, respectively.

5 Optimization of Source/Drain Resistance

5.1 Calculation of External Resistance

From the conventional drift-diffusion theory of electron transport in semiconductors, the sheet resistivity (R_{sh}) is calculated from an incremental form of Ohm's law [11]:

$$I_{ds} = \int_0^\infty J(y)dy = \frac{d}{dx}\phi_n(x)\int_0^\infty qn(x,y)\mu(x,y)dy = \frac{1}{R_{sh}(x)}\frac{d}{dx}\phi_n(x) \tag{4}$$

where x is the direction parallel to the Si surface, y is the direction perpendicular to the interface, and ϕ_n is the electron quasi-Fermi level obtained from 2D device simulation. I_{ds} is independent of x due to current continuity. The sheet resistance is simply proportional to the lateral gradient of $\phi_n(x)$. This expression is valid as long as the device is sufficiently wide and the current flow is largely parallel to the $x-$direction. Also, the total resistance (R_{tot}) is calculated by the integration of $R_{sh}(x)$ along the direction from source to drain. Fig. 8 (a) shows the external resistance components in a conventional HDD NMOS transistor: (i) the resistance in the contact (R_{co}), (ii) the sheet resistance of the deep source/drain junction (R_{sh}), (iii) the spreading resistance (R_{sp}) of the SDE, (iv) the accumulation-layer resistance under the gate overlap (R_{acc}). While the accumulation layer in the HDD region is similar in many respects that of the inversion layer in the channel, certain fundamental differences exist because of the nature of scattering mechanisms as carriers enter the accumulation region. Hence, a unified mobility model for the inversion and accumulation layers is needed for precise resistance calculations.

5.2 Mobility Model for Inversion/Accumulation Layer

Due to the two-dimensional nature of electron gas in the accumulation layer, mobility degradation occurs in a manner similar to that in the inversion layer owing to the transverse electric field. For an accurate accumulation resistance (R_{acc}) calculation, however, the mobility model should account for the screening of Coulombic scattering since the electron concentrations in the accumulation layer exceed the donor concentration [12]. The principal difference between Coulombic scattering in the accumulation and inversion layers is that electrons scatter off positively charged donor atoms in an accumulation layer, whereas electrons scatter off negatively charged acceptor atoms. Thus, Coulombic scattering is stronger in the accumulation layer compared to the inversion layer. Without proper mobility modeling, lower external resistance is predicted, resulting in an under-estimation of degradation due to R_{acc}.

In this work, the mobility model considers both unscreened and screened Coulombic contributions as well as the phonon and surface roughness scattering terms. Fig. 8 (b) shows measured and simulated total resistance (R_{tot}) using the mobility model as a function of the V_{gs} for the NMOS with the $L_{eff} = 0.25\ \mu$m, which was fabricated in the Hewlett-Packard Laboratories.

With the process and device models used in simulation of an $L_{eff} = 0.25\ \mu$m experimental transistor, a 0.08 μm L_{eff} NMOS simulated structure was created using 2D process simulation to investigate the external resistance effects for sub-100 μm MOSFET's. The simulated shallow SDE uses an arsenic implantation with dose of $1 \times 10^{15}\,\mathrm{cm}^{-2}$, and energy in the range of 5-30 KeV, producing SDE depths range from 30 to 70 nm.

5.3 Shallow S/D Extension and Gate Overlap Effects

Fig. 9 (a) shows the computed sheet resistance along the channel direction for different V_{gs} biases in the linear operation region (i.e. $V_{ds} = 0.05\,\mathrm{V}$). It should be noted that the sheet resistance in the channel and accumulation regions (outside the metallurgical channel) is highly dependent upon the gate bias, and the peak sheet resistance occurs in the accumulation region under the gate overlap (L_{ov}) for $V_{gs} > V_{TH} = 0.45\,\mathrm{V}$. This gate modulated series resistance is associated with the laterally finite source/drain doping gradient. Because of this finite lateral gradient, current injection from the surface inversion layer into the source/drain does not occur immediately at the metallurgical junction. This implies that a significant voltage drop occurs in the extrinsic region. Namely, the sheet resistance in the accumulation region and current crowding/spreading region is comparable to the sheet resistance in the inversion layer in sub-100 nm MOSFET's. The contact resistance (R_{co}) for simulation is assumed 4 $\Omega/\square$, which is a typical value of silicide processes.

Fig. 9 (b) shows R_{ext} with respect to the gate bias for two gate oxide thicknesses (2 and 4 nm). Quantum mechanical effects due to the thin gate oxides are ignored in this work. The accumulation resistance (R_{acc}) decreases as the gate oxide thickness is reduced, because R_{acc} is inversely proportional to t_{ox}. In addition, R_{ext} abruptly increases as the SDE depth (X_j) becomes shallower than 50 nm, as shown in Fig. 10 (a). Also, R_{ext} increases sharply when L_{ov} is shorter than 20 nm because the gate coupling to SDE becomes poor, as shown in Fig. 10 (b). Thus, careful optimization of X_j and L_{ov} is necessary to minimize the R_{ext}. Otherwise, I_{dsat} and SDE to gate coupling can be degraded even though the short channel effects are reduced. The R_{sh} distributions in the various X_j at the saturation region of $V_{ds} = V_{gs} = 1.5\,\mathrm{V}$ are shown in Fig. 11 (a), where the maximum resistance appears in the drain accumulation region. Here, R_d increases as the X_j becomes shallower, which implies I_{dsat} degradation for the SDE depths less than 40 nm. In order to observe the R_{ext} effects on I_{DSAT}, simulations were performed for the different X_j's as in Fig. 11 (b). To consider the short channel effects (i.e., DIBL) induced leakage current, the V_{dsat} for each X_j condition was chosen as the drain bias that produces 1 nA/μm of off-state leakage current. As a result, I_{dsat} degradation is observed for the X_j less than 40 nm, which corresponds to the experimental trend in [2].

6 Compact Gate Tunneling Model for Circuit Simulation

A direct tunneling current in the NMOS structure as in Fig. 12 is expressed as [13],

$$J_{DT} = C(V_{ox}/t_{ox})^2 e^{\frac{-B(1-(1-V_{ox}/\phi_B)^{3/2})}{V_{ox}/t_{ox}}} \tag{5}$$

where B and C are physical parameters (see [13]). In order to use Eq. (5), it is necessary to relate the oxide voltage (V_{ox}) to the applied voltage (V_{gs}), since V_{ox} depends on V_{gs} as well as the surface potential (ψ_s) as:

$$V_{ox} = V_{gs} - V_{FB} - \psi_s \tag{6}$$

However, ψ_s can only be solved accurately using an iterative numerical approach. A compact model with an explicit formulation of ψ_s [14] is used in this work such that in the weak inversion region ($0 < \psi_s < 2\phi_f$) the surface potential can be approximated as

$$\psi_{s,weak} = V_{gb} - V_{FB} + \frac{\gamma^2}{2} - \gamma\sqrt{V_{gb} - V_{FB} + \frac{\gamma^2}{4}} \tag{7}$$

where γ is the body factor defined by $\sqrt{2q\epsilon_{Si}N_a}/C_{ox}$. In the strong inversion region ($\psi_s > 2\phi_f$) ψ_s becomes

$$\psi_{s,strong} = 2\phi_f + V \tag{8}$$

where V denotes the electron quasi-Fermi potential, ranging from V_{sb} at the source and to V_{db} at the drain side. The surface potential at the drain node ($\psi_{s,drain}$) is modeled to consider the drain bias (V_{ds}) effect as

$$\psi_{s,drain} = \psi_s(V_{ds} + V_{sb}), \quad \psi_{s,src} = \psi_s(V_{sb}) \tag{9}$$

In addition, ΔV_{gs} is modeled and added to V_{gs} to consider the channel length dependence on gate current related to DIBL as

$$\Delta V_{GS} = \delta_{DIBL}\sqrt{2\phi_f + V_{SB}} \cdot V_{DS} \tag{10}$$

where $\delta_{DIBL} = \delta_O(\frac{L}{L_R})^{n_{dibl}}$, L_R is the reference gate length and n_{dibl} is an exponent of the length dependence. For an exact calculation of surface potential considering quantum-mechanical effects, an approximate manner by using van Dort's bandgap broadening approach is used

$$\Delta E_g = \frac{\beta}{2kT}E_n^{2/3} \tag{11}$$

where E_n is the normal electric field at the Si-SiO$_2$ interface which is gate bias dependent and β is a factor that can be determined experimentally. The value of ΔE_g is used to calculate a new intrinsic carrier concentration n_i and ϕ_f.

In the classical case, the electron density has its maximum value at the Si-SiO$_2$ interface, while in the quantum mechanical case the electron density is diminished at the interface, increases to its maximum value and decreases with the distance from the surface. To model this, the surface electron concentration n_s in the inversion region is expressed as,

$$n_s = [1 - e^{-z_0^2/\lambda_{th}^2}]N_{sub}e^{(\psi_s - 2\psi_f)/kT} \tag{12}$$

where λ_{th} is the thermal wavelength as determined by the carrier effective mass with theoretical value of electron being 1.27 nm and z_0 is introduced for the finite concentration at the interface ($z = 0$).

Fig. 13 and 14 shows V_{ox} and simulated gate currents using the new model, the conventional methods [13] and 1D Green's function solver, NEMO. Comparing NEMO and real measurements, simulated gate currents for different oxide thicknesses are shown in Fig. 15 and 16. The results between the compact model and numerical solver agree very well due to the considerations of the surface potential as well as quantum mechanical effects. The simulated gate tunneling current decreases as the drain bias increases since the potential difference between the gate and the channel surface becomes smaller, as shown in Fig. 17. Figs. 18 and 19 illustrate the channel-length dependence on gate and drain currents; I_{gate} decreases in inverse proportion to $L_g^{1.5}$, which is comparable to the experiments in the reference [10]. Transient circuit simulation using this model is performed as shown in Fig. 20, where the tunneling current effect is remarkable, even when the gate bias is low.

7 Summary

MOS C–V characteristics of gate oxide in the sub-2.0 nm region are modeled with an empirical, hybrid formulation for QM corrections implemented in a 2D device simulator. The sharp decrease in capacitance for gate oxides below 2.0 nm in both the accumulation and inversion regions is modeled by using a distributed RC network that includes the gate tunneling current which is calculated using a Green's function solver. Also, an accurate direct tunneling model for circuit simulation is modeled by incorporation of the explicit surface potential model and quantum-mechanical corrections. In addition, accurate determination of external resistance in scaled MOS can be addressed using proper mobility modeling. External resistance in the accumulation layer (R_{acc}) becomes dominant for the shallow SDE depth less than 40 nm. R_{acc} is highly dependent on the junction depth and overlap length, which should be minimized for improvement of current driving capability and SDE to gate coupling contrary to the future trends in using elevated source & drain structure [15]. An nMOS transistor with L_{eff} of 0.08 μm shows optimum I_{dsat} when L_{ov} is 20 nm and X_j is 40 nm.

Acknowlegement: The support from NSF throught NCCE and DesCArtES projects is greatly appreciated.

References

[1] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S. H. Lo, and *et al.*, "CMOS scaling into nanometer regime", *IEEE Proceedings*, vol. 85, no. 4, p. 486, 1997.

[2] S. Thompson, P. Packan, M. Alavi, I. Post, S. Tyagi, S. Ahmed, S. Yang, and M. Bohr, "Source/Drain Extension Scaling for 0.1μm and Below Channel Length MOSFETs", *IEEE Symp. on VLSI Technology Digest*, pp.132–133, 1998.

[3] W.K. Henson, K.Z. Ahmed, E.M. Vogel, J.R. Hauser, J.J. Wortman, R.D. Venables, M. Xu, and D. Venables, "Estimating Oxide Thickness of Tunnel Oxides Down to 1.4 nm Using Conventional Capacitance-Voltage Measurements on MOS Capacitors", *IEEE Electron Device Letters*, vol. 20, pp. 179–181, Apr. 1999.

[4] S. Biesemans, M. Hendricks, S. Kubicek, and K. De Meyer, "Practical Accuracy Analysis of Some Existing Effective Channel Length and Series Resistance Extraction Methods for MOSFET's", *IEEE Tran. on Electron Devices*, vol. 45, no. 6, pp. 1310–1316, June, 1998.

[5] C.H. Choi, J.S. Goo, T. Y. Oh, Z. Yu, R. W. Dutton, A. Bayoumi, M. Cao, P. V. Voorde, D. Vook, and C. H. Diaz, "MOS C–V Characterization of Ultra-Thin Gate Oxide Thickness (1.3–1.8 nm)", *IEEE Electron Device Letters*, vol. 20, pp. 292–294, June 1999.

[6] P. Vande Voorde, P. B. Griffin, Z. Yu, S-Y Oh, and R. W. Dutton, "Accurate doping profile determination using TED/QM models extensible to sub-quarter micron nMOSFETs", *Digest IEDM* '96, p. 811, San Francisco, 1996.

[7] C. Bowen, C. Fernando, G. Klimeck, Enda, and *et al.*, "Physical oxide thickness extraction and verification using quantum mechanical simulation", *Digest IEDM*, p. 869, 1997.

[8] S. Nagano, and M. Tsukiji, E. Hasegawa, and A. Ishitani, "Mechanism of leakage current through the nano-scale SiO_2 layer", *J. of Applied Physics*, vol. 75, no. 7, p. 3530, 1994.

[9] C.H. Choi, J.S. Goo, T. Y. Oh, Z. Yu, R. W. Dutton, A. Bayoumi, M. Cao, P. V. Voorde, and D. Vook, "C–V and Gate Tunneling Current Characterization of Ultra-Thin Gate Oxide MOS (t_{ox}=1.3–1.8 nm)", *Symposium on VLSI Technology*, pp. 151-152, 1999.

[10] H.S. Momose, M. Ono, T. Yoshitomo, T. Ohguro, S. Nakamura, M. Saito, and H. Iwai, "1.5 nm Direct-Tunneling Gate Oxide Si MOSFET's", *IEEE Tran. on Electron Devices*, vol. 43, no. 8, pp.1233–1242, Aug. 1996.

[11] S. A. Mujtaba, S. Takagi, R. W. Dutton, "Accurate Modeling of Coulombic Scattering, and Its Impact on Scaled MOSFETs," *IEEE Symp. on VLSI Technology Digest*, pp.99–100, 1995.

[12] Y. Taur and T. Ning, *Fundamental of modern VLSI device*, Cambridge University Press, 1998.

[13] K.F. Schuegraf and C. Hu, "Hole Injection SiO_2 Breakdown Model for Very Low Voltage Lifetime Extrapolation", *IEEE Trac. on Electron Devices*, vol.41, p.761, 1994.

[14] R.van Langevelda, *PhD thesis, Eindhoven Univ. of Tech.*, 1998.

[15] H. Iwai, "CMOS Technology – Year 2010 and Beyond", *IEEE J. of Solid-State Circuits*, pp. 357–366, vol. 34, no. 3, March 1999.