Predictive Compact Modeling for Strain Effects in Nanoscale Transistors

Nuo Xu*, Xin Sun, Lynn Wang, Andrew Neureuther, and Tsu-Jae King Liu
Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720 USA
Phone: +1-510-643-2639, Fax: +1-510-643-2636, *E-mail: nuoxu@eecs.berkeley.edu

***Abstract*— A compact MOSFET I-V model is developed based on quasi-ballistic transport theory, using a more accurate method to calculate the effective stress and its impact on all strain-dependent parameters. This model is verified using published 40nm-L_g CESL-strained nMOSFET data, and can be used to predict layout-dependent variations and future-generation device performance trends.**

Keywords- critical length; strain effects; backscattering rates; quasi-ballistic transport; predictive modeling

I. INTRODUCTION

As the gate length (L_g) of a MOSFET is scaled down and carrier transport is enhanced increasingly via process-induced strain, the transistor drive current becomes limited by the carrier injection velocity from the source into the channel [1]. Hence, the quasi-ballistic transport model proposed by Lundstrom *et al.* [2] is more suitable for predicting nanoscale MOSFET behavior. By self-consistently solving for the charge density at the top of the potential barrier and the surface Fermi level, it is straightforward to calculate the transistor current:

$$I = \frac{qW\sqrt{2m_y}\left(\frac{k_BT}{\pi}\right)^{\frac{3}{2}}}{2\hbar^2}(1-R)[\Im_{\frac{1}{2}}(E_{f1}) - \Im_{\frac{1}{2}}(E_{f2})] \quad (1)$$

The top-of-barrier ("critical region") length l is dependent on the applied drain bias, and can be approximated by assuming a parabolic potential profile along the channel [3], as verified by parameter extractions from experimental data [4]. l can be expressed as follows:

$$l = L_g \times \sqrt{\frac{k_BT}{qV_d}} \times tan^{-1}\left(\sqrt{\frac{qV_d}{k_BT}}\right) \quad (2)$$

The effect of a small number of carrier-scattering events is taken into account by including a backscattering rate parameter R [2]:

$$R = \frac{l}{l + \lambda_0} \quad (3)$$

In previous transistor models based on injection velocity and quasi-ballistic transport, the effective stress was simply taken to be the average value of stress along the entire length of the channel, and only sub-band energy splitting induced by strain was taken into account [5]. We present here an improved model that more accurately calculates the effective stress which affects the carrier injection velocity, and that accounts for all strain-dependent model parameters.

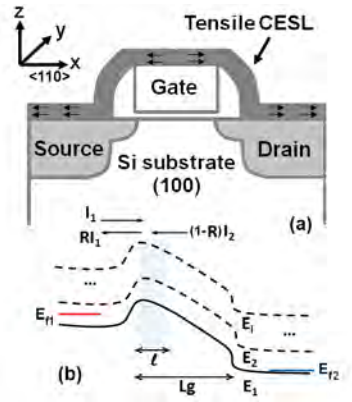

Figure 1. Schematic view of strained nMOSFET with tensile CESL layer (a) and quasi-ballistic transport model (b).

II. MODEL FOR EFFECTIVE STRESS AND STRAIN EFFECTS

A. *Effective Stress Calculation*

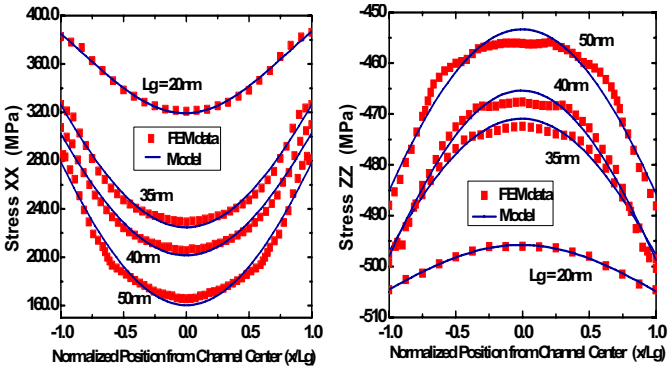

Figure 2. Simulated stress XX (left) and ZZ (right) distributions for various channel lengths, obtained using Sentaurus (Synopsys, Inc.).

A tensile contact etch-stop liner (CESL) encapsulating a MOSFET (Fig. 1) induces local non-uniform tensile and compressive stress within the channel in the *x*- and *z*-directions, respectively. Fig. 2 shows simulated stress profiles in the MOSFET channel for a 60nm-thick and 1GPa (tensile) CESL. It is clearly seen that the stress components vary significantly with position along the length of the channel and with L_g in the range from 20nm to 50nm. The stress profiles can be well modeled as Gaussian functions, with fitting parameters L_0 and $A_{1\text{-}3}$ that depend on the process technology:

$$S(x, L_g) = (L_g - L_0)\left\{A_1 exp\left[-\frac{1}{2}\left(\frac{x}{L_g - L_0}\right)^2\right] + A_2 \exp\left(-\frac{L_g}{L_0}\right) + A_3\right\} \quad (4)$$

The value of effective stress that should be used to calculate the carrier injection velocity is obtained by integrating the stress in the critical region and dividing by the critical length *l*. Since the integral of a Gaussian function is an analytical function, the effective stress can be expressed in compact form:

$$\bar{s}(l,L_g)=(L_g-L_0)\left[A_2\exp\left(-\frac{L_g}{L_0}\right)+A_3\right]+\frac{\sqrt{2\pi}A_1(L_g-L_0)^2}{2l}\left\{\mathrm{erf}\left[\frac{2l-L_g}{2\sqrt{2}(L_g-L_0)}\right]+\mathrm{erf}\left[\frac{L_g}{2\sqrt{2}(L_g-L_0)}\right]\right\} \quad (5)$$

Fig. 3 shows how *l* and the effective stress in the *x*-(channel) direction depend on the applied drain bias, for L_g=40nm. The non-monotonic dependence of the effective stress on the drain bias can be attributed to the Gaussian stress profile (ref. Fig. 2).

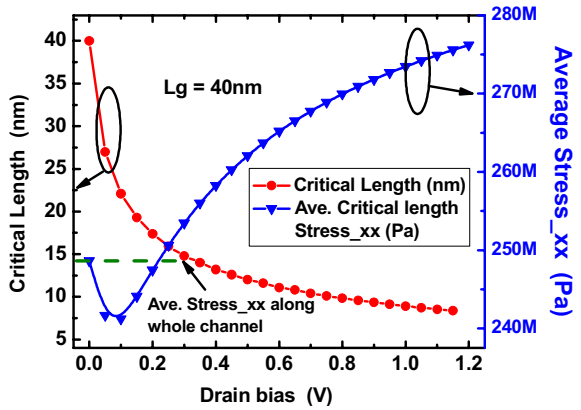

Figure 3. Critical length and effective stress versus drain bias.

B. *Modeling Quantization Effects in Bulk nMOSFET*

Our transistor model is based on quasi-ballistic transport theory. The improved Airy's approach proposed in [6] is used to analytically calculate the sub-band energy level, where the effective field versus charge dependence factor f is fitted for different sub-bands, separately:

$$\mathrm{E_i}^{l,t}=\left(\frac{q\hbar}{\sqrt{2m^{l,t}}}F_s\right)^{\frac{2}{3}}\times AiryRoots(i) \quad (6)$$

$$F_s=q\frac{(N_{dep}+fN_s)}{\varepsilon_{Si}} \quad (7)$$

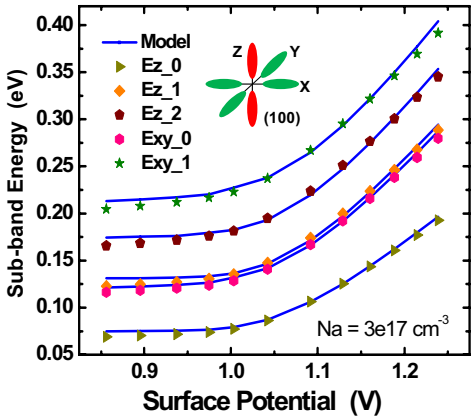

Figure 4. Sub-band energy levels versus (100) inversion-layer potential. The analytical model is compared against Poisson-Schrodinger results (w/o strain).

Fig. 4 compares the modeled nMOSFET sub-band energies against those obtained by self-consistently solving the Poisson and Schrodinger equations. Since numerical simulations indicate that the five lowest sub-bands of the conduction band (3 in z-valleys and 2 in x,y-valleys, for Si with (100) surface orientation) hold over 99% of the electrons over the range of typical gate biases, we only consider electron density and conduction in these 5 sub-bands. The quantization-induced sub-band energy shifts are added to the strain-induced band energy splitting to account for carrier redistribution among the different sub-bands.

C. *Physical Modeling of Strain Effects*

In order to calculate the backscattering rate, the mean free path (MFP) for degenerate carriers under low-field (Eqn. 8) is needed and thus the carrier mobility first must be calculated.

$$\lambda=\left(\frac{2\mu}{\nu_T}\frac{k_BT}{q}\right)\frac{\Im_0^2(E_f)}{\Im_1'(E_f)\times\Im_{\frac{1}{2}}(E_f)} \quad (8)$$

Compact models traditionally use piezo-resistance theory to describe the effect of strain on bulk carrier mobility. This approach is inaccurate for predicting mobility enhancement at high levels of strain, however, because it eventually saturates. The analytical model for bulk electron mobility suggested by Dhar *et al.* [7,9] is used in our model to capture this behavior. It accounts for sub-band splitting and reduction in inter-valley scattering. Relevant parameters (*e.g.* phonon energy and coupling constants) are extracted and better fit the reported experimental data for biaxially tensile-strained Si (Fig. 5) [8]. In our model, the effective low-field mobility is calculated as a weight average, considering the different occupation rates among different sub-bands, for each value of applied gate bias.

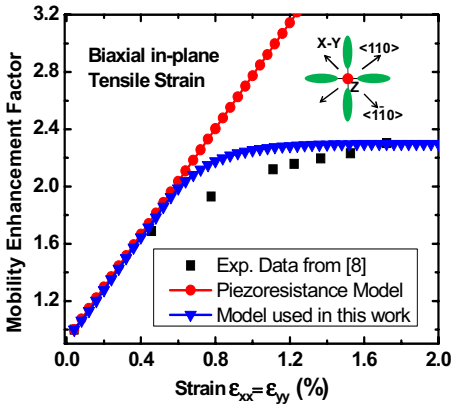

Figure 5. Electron mobility as function of strain using the piezo-resistance model versus the scattering model proposed in [7,9].

Under <110>/(100) uniaxial stress, the z-valleys are warped due to the non-parabolic portion of the conduction band and thus corrections to both the transverse (parallel and vertical to the channel direction) and longitudinal effective masses are needed. We use the effective mass model in [10], which compares well with numerical non-local Empirical Pseudopotential Method (EPM) calculations. Transverse mass variation in the z-valleys will affect channel conductivity effective mass and density of states, while longitudinal mass

changes will affect the quantization energy. For the case of an nMOSFET with L_g=40nm, these effective-mass corrections result in 3%~14% increase in transistor drive current gain across the range of operating voltages (Fig. 6).

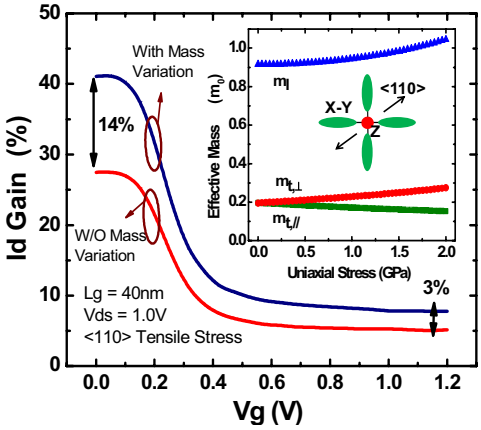

Figure 6. Transistor drive current gain induced by <110>/(100) uniaxial stress. The insert shows m_t and m_l changes vs. strain.

Fig. 7 illustrates the procedure used to calculate the MOSFET current in our model. First, for a given drain bias, the critical-region length is calculated. Next, the effective stress (*i.e.*, the average stress value in the critical region) is calculated and used to correct the band deformation energy and calculate the effective mass variation. Then, for each value of applied gate bias, an iterative process is used to self-consistently solve for the sub-band energy levels, surface Fermi level, and electron density with the back-scattering rates (calculated using the low-field mobility value). Finally, the drain current is calculated to be the sum of the sub-band currents for the 6 valleys.

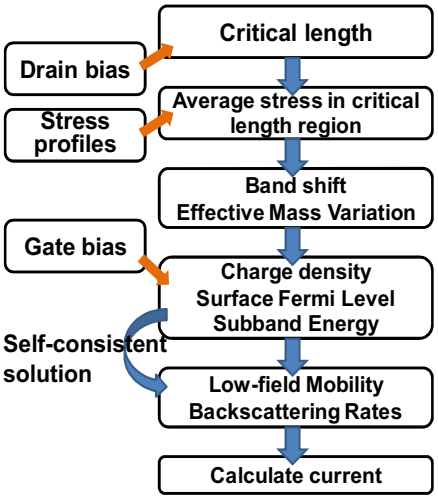

Figure 7. Procedure used for the physically based compact I-V model for nanoscale MOSFETs.

III. Model Verification and Predictions

Our model is validated using experimental data for 40nm-L_g nMOSFETs with tensile CESL reported in [11]. The stress distribution in the channel is taken from published simulation results and calibrated with current gain *vs.* channel length data for L_g=40nm, in order to calculate the effective stress. An effective channel doping level of 10^{18} cm^{-3} is extracted from the sub-threshold swing value at low drain bias (V_{ds} = 0.05V). Fig. 8 shows that a good fit to the experimental data is obtained with our model. Note that the only fitting parameters are gate/drain control coefficients, low-field mobility, and source/drain series resistance.

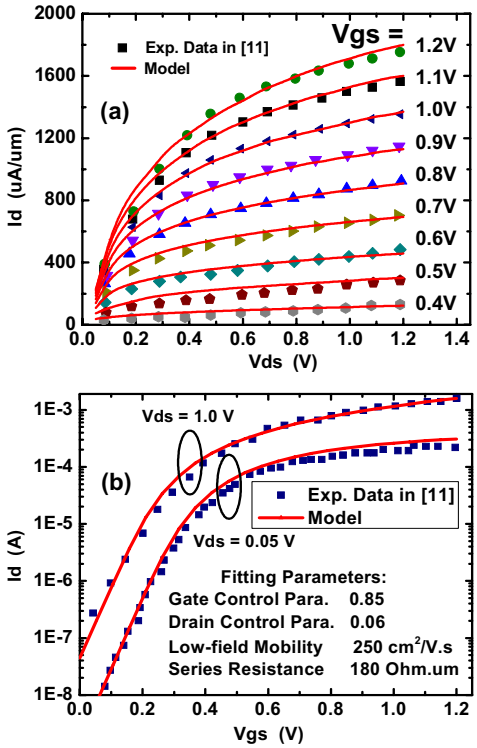

Figure 8. Comparison of modeled *vs.* measured output (a) and transfer (b) characteristics for nMOSFET with Lg=40nm, Tinv =1.4nm.

With parameters fitted to L_g=40nm nMOSFET I-V data, our model accurately predicts trends in MOSFET performance for L_g ranging from 20nm to 100nm (Fig. 9), for large drain bias. Since it is physically based, it should more accurately predict the effects of stress-induced layout dependent variations, as compared with empirical compact models. The dependencies of stress profiles on other layout parameters (*e.g.* channel width, active area length, isolation length, *etc.*) can be easily included in the analytical stress model.

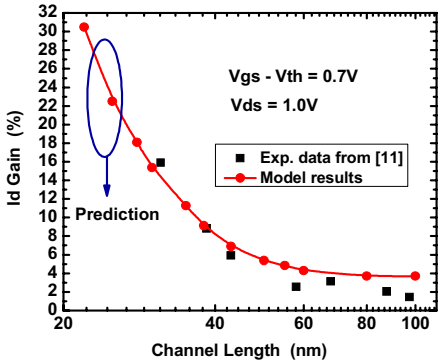

Figure 9. Simulated dependence of transistor drive current enhancement on channel length, for large drain bias.

Our model also can be used to predict the performance of future-generation devices, considering the projected MOSFET parameter values for the HP90 to HP40 nodes in the ITRS [12], listed in Table I. The channel doping is assumed to be optimized to maintain constant drain-induced barrier lowering (DIBL) and body effects (*i.e.* constant gate and drain control coefficients), and the mobility degradation under transverse electric field is interpolated from data in [13]. The stress liner thickness and initial stress value is assumed to remain the same as for the L_g=40nm technology in [11], for simplicity.

TABLE I. DEVICE PARAMETERS FOR FUTURE TECHNOLOGY NODES, FROM ITRS.

HP Tech Node	*90nm*	*68nm*	*52nm*	*40nm*
L_g (nm)	32	25	20	16
T_{inv}(nm)	1.93	1.84	1.04	0.82
Body Doping (cm^{-3})	3.3e18	4.8e18	4.1e18	6.6e18
S/D Series Resistance (Ohm.um)	180	200	200	180
Vdd (V)	1.1	1.1	1	1

Fig. 10 shows the predicted injection velocity *vs.* gate length from 32nm to 16nm. The results show that the enhancement in drive current will saturate as L_g scales down, partly due to the fact that source/drain series resistance is not expected to scale down. The ballistic velocity (without strain) is also shown for reference in Fig. 10. As expected, modern transistor performance is far from the ballistic limit, even with uniaxial strain technology [1].

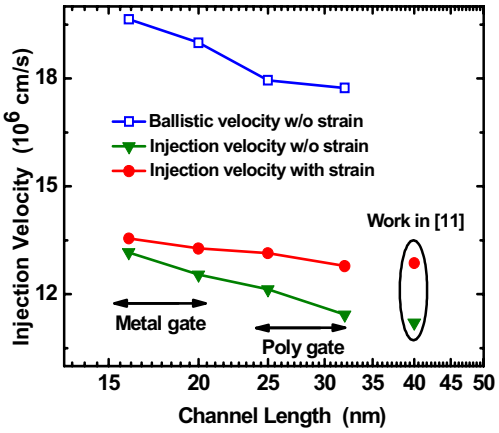

Figure 10. Model prediction of injection velocity vs. channel length.

Fig. 11 shows how the intrinsic delay (calculated using a more realistic model than the conventional CV/I formulation [1]) is predicted to improve with L_g scaling. These results show that the ITRS is optimistic in estimating delay, and also that strain technology will have decreasing benefit as the impact of parasitic capacitance (*e.g.* fringing capacitance) increases with channel length scaling.

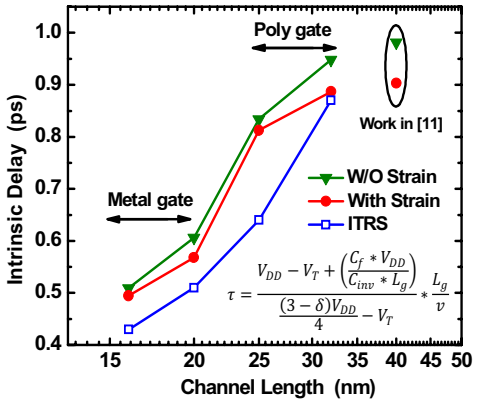

Figure 11. Model prediction of intrinsic transistor delay vs. channel length.

IV. SUMMARY

A physically based transistor model with only 4 fitting parameters is developed based on injection velocity and quasi-ballistic transport theory, and is validated using published experimental data for a L_g=40nm CESL-strained nMOSFET. This compact model can be used to predict layout-dependent transistor variations due to local non-uniform stress, as well as future-generation device performance trends.

REFERENCES

[1] A. Khakifirooz *et al. IEDM Tech Dig.*, p.667 (2006).
[2] A. Rahman *et al. IEEE-TED* vol.49, p.481 (2002).
[3] R. Kim *et al. IEEE-TED* vol.56, p.132 (2009).
[4] M. J. Chen *et al. IEEE-EDL* vol.28, p.177 (2007).
[5] E. Fuchs *et al. SISPAD Proc.*, p.303 (2005).
[6] M. Ferrier *et al. SSE* vol.50, p.69 (2006).
[7] S. Dhar *et al. IEEE-TED* vol.52, p.527 (2005).
[8] K. Rim *et al. IEDM Tech Dig.*, p.50 (2003).
[9] S. Dhar *et al. IEEE-Trans. Nano*. vol.6, p.97 (2007).
[10] E. Ungerboeck *et al. IEEE-TED* vol.54, p.2183 (2007).
[11] S. Mayuzumi *et al. IEDM Tech Dig.*, p.293 (2007).
[12] ITRS 2007 Edition.
[13] S. Takagi *et al. IEEE-TED* vol.41, p.2357 (1994).