MOSFET Modeling Into the Ballistic Regime

J.D. Bude

Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974

Abstract – Physically-based full band Monte-Carlo simulations are compared with drift-diffusion simulations for channel lengths from 150nm to 40nm. Errors in the drift-diffusion simulated I_{ON}, g_m and channel velocities are quantified through comparison with Monte-Carlo simulations under realistic surface scattering conditions. Suggestions for improving the drift-diffusion results are also discussed.

1. Introduction

Based on the first moment of the Boltzmann equation, drift-diffusion (DD) is only strictly valid for the low field near- equilibrium conditions found in long channel MOSFETs. However, despite its well-documented inadequacies, it has been found to be surprisingly accurate for channels lengths (L_{CH}) as short as 0.1μm [1]. Although full CMOS processes with effective channel lengths (L_{eff}) < 90nm [2]-[3] and experimental nMOS and pMOS down to 25nm which exhibit near ballistic transport [4] have recently become available, no systematic assessment of the applicability of the DD equations at these extremes exists [5].

It is the purpose of this paper to clarify the accuracy of DD models as L_{eff} is scaled as short as 30nm by means of comparison to self-consistent Monte Carlo (MC) simulations. The MC model has been carefully calibrated to yield bulk field dependent velocities which are nearly identical to those used in the DD model, thereby directly highlighting the errors in the DD transport model. Realistic values for the inversion layer carrier velocities were used. Section 2 discusses the simulation methodology, sections 3 and 4 present results for nMOS and pMOS and section 5 suggests possible improvements for drift-diffusion in the near-ballistic regime.

2. Simulation Methodology

For the purposes of comparing DD and MC, a simple planar MOSFET structure with constant channel doping and an oxide thickness of 2.2nm (physical thickness ~ 1.5nm) was adopted. Gate lengths (L_{CH}) from 40nm to 150nm were simulated, and the channel doping was adjusted to yield threshold voltages of about 0.4V for nMOS and 0.3V for pMOS at V_D=1V with L_{CH} independent I_{OFF} of about 1nA/μm. The shallow junction lateral diffusion length was about 5nm so that $L_{eff} \approx L_{CH}$ – 10 nm. DD simulations employed the Caughey-Thomas (CT) field dependent velocity model [6]: $v(E) = E\,\mu(E) = \mu_0 / (1 + (E\mu_0/v_{sat})^{\beta})^{1/\beta}$. Here, E is the lateral electric field (parallel to the oxide interface), v_{sat} is the saturation velocity, μ_0 is the low field inversion layer mobility and β is a constant. Measured values for electrons are v_{sat}=9.5 10^6 cm/s and β = 2 [7], and these values are used in sections 2 and 3. There is a spread in reported values for the hole v_{sat}: values as high as 8.5 10^6 cm/s have been used in DD simulators, but MC simulations typically give a value closer to 6 10^6 cm/s [8] and inversion layer saturation velocities < 5 10^6 cm/s have been measured [9]. In order to make direct comparison to MC, an average value of v_{sat} for holes was taken as 6.5e6 cm/s; the measured value of β=1 was used for holes. These values were used in sections 2 and 4. In the following, μ_0 was taken as a constant in the device to remove any differences in the inversion layer values calculated with MC or DD. Both MC and DD simulations used Boltzmann statistics.

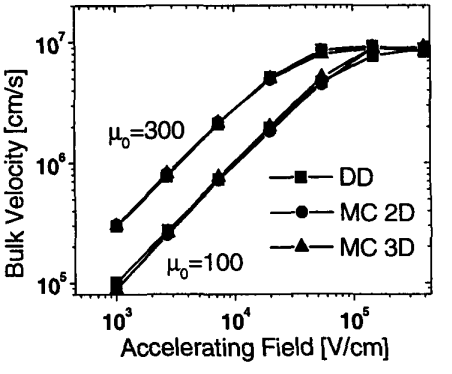

Fig. 1: Comparison of electron field dependent velocity (constant field) with μ_0=300 cm^2 /Vs and 100 cm^2 / Vs.

Self-consistent MC simulations were performed using the full-band simulator SMC [10]. Most MC models treat inversion layer μ_0 reduction by adjusting the fraction of diffusive scattering from the oxide interface [11]. To ensure that MC exactly reproduced DD values for μ(E) in long channel devices, a position dependent low energy scattering rate (physically motivated) was added to the bulk scattering rates (no diffusive interface scattering) to give the correct μ(E).

Also, for electrons in the presence of strong confining fields, mainly the low mass X-valleys participate in transport. To include this effect simply in the MC, the probability of a carrier scattering into a higher-mass X-valley was adjusted based on electron energy and the value of the confining field. Again, μ(E) was computed in long channel devices and fit to the DD model. Device results with this quasi-2D effect are labeled "2D" and transport including all X-valleys is labeled "3D". Fig. 1 shows the computed bulk

0-7803-6279-9/00/$10.00 © 2000 IEEE

electron velocity for the MC models and the CT expression used in the DD simulations; less than a 3% difference in velocity occurs for E<100kV/cm. Less than a 5% difference is found comparing MC and DD hole velocity.

3. nMOSFET Results

Fig. 2 shows the electron velocity in the channel for the L_{CH}=40nm and 150nm devices with μ_0=300 cm^2/Vs (a reasonable value for these effective fields [12]). Strong velocity overshoot is seen in the MC simulations. However, the injection velocity (v_{IN}) at the top of the source potential barrier (X_S) which controls the channel current flow is similar in both simulations (X_S is the point where the lateral electric field passes through zero). In the 40nm device the velocity distribution at X_S is nearly ballistic (78% transmission probability [4]).

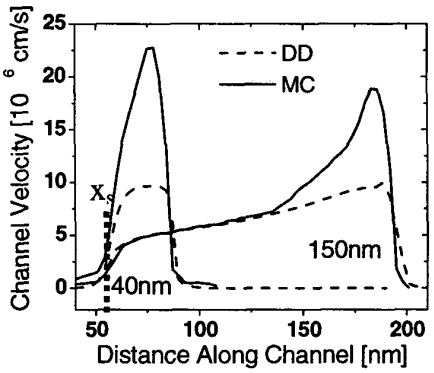

Fig. 2: Channel velocity in L_{CH} = 40nm and 150nm devices averaged perpendicularly from the oxide interface through 4nm of the channel.

Although all the assumptions in the DD transport model are violated in this extreme, the similarity of v_{TH} and v_{sat} [13] and the form of the CT velocity model yields a v_{IN} underestimate of only 13% (fig. 3).

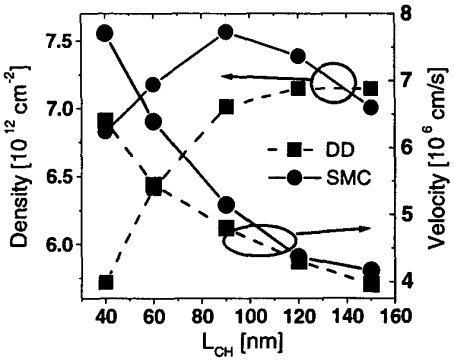

Fig. 3: n_{IN} and v_{IN} at the top of the source barrier as a function of L_{CH} with V_{GS}=V_{DS}=1V. Averages as in fig. 3. μ_0 = 300 cm2 / Vs. Note, I_D=q n_{IN} v_{IN}.

However, DD also underestimates carrier density (n_{IN}) at the top of the barrier, and the drain current (I_D) is determined by both: I_D=q * v_{IN} * n_{IN}. Fig. 3 shows n_{IN} and v_{IN} as a function of L_{CH}. As the channel is shortened, DD n_{IN} decreases sharply due to 2D charge sharing and the increased channel doping needed to control short channel effects. MC n_{IN} does not decrease drastically due to an additional barrier lowering effect. The channel velocity overshoot leads to a non-uniform reduction in the channel carrier density with respect to DD, and the surface potential must be pulled down to accommodate it.

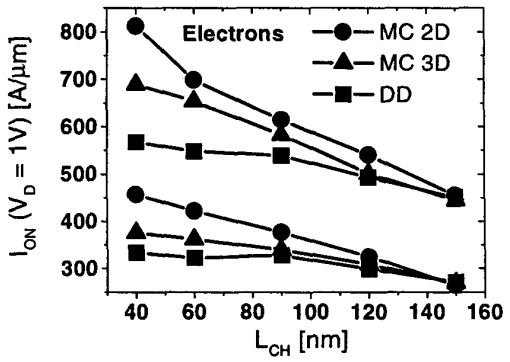

Fig. 4: nMOS I_{ON} as a function of L_{CH} with μ_0 = 300 cm^2 / Vs and V_{GS}=0.8V and 1V.

This effect has been seen qualitatively in energy transport simulations of velocity overshoot [14]. Here, it is accounted for rigorously. In the 150nm device, most of the reduction occurs near the drain and the effect on the source barrier is small. However, in the 40nm device, the proximity of the velocity overshoot to the source results in a substantial barrier lowering with respect to the DD simulation and a corresponding increase in n_{IN}. In addition to barrier lowering, velocity overshoot also reduces small signal gate capacitance.

These trends are clearly illustrated in simulated I_{ON} (fig. 4). While DD makes little error in the longer channel devices, it underestimates I_{ON} by 40% due to the combined effects of increased MC v_{IN} (13% larger) and n_{IN} (23% larger). Fig. 5 shows the error in g_m is even larger. As a result, DD underestimates the scaling trends: DD suggests that scaling from 150nm to 40nm while maintaining constant I_{OFF} only results in a 26% increase in I_{ON}; MC predicts an 80% increase.

The change in the transport mass due to quantization becomes important below 80nm. Despite having the same long channel velocity behavior, as L_{CH} is reduced towards the ballistic regime the importance of the mass through an increased v_{TH} is highlighted.

Fig. 6 shows the error in I_{ON} as a function of the inversion layer mobility. As μ_0 is reduced, the assumptions in DD become more reasonable, and the difference with MC is reduced. Fig. 6 also shows that the impact on performance when using a high K dielectric (where the mobility may be less than SiO_2) or operating at higher inversion fields is larger than expected from DD models.

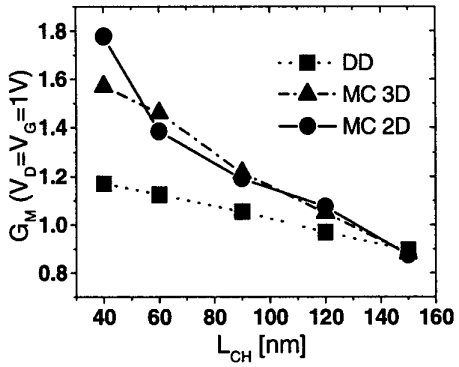

Fig. 5: transconductance S/mm for devices of fig. 4.

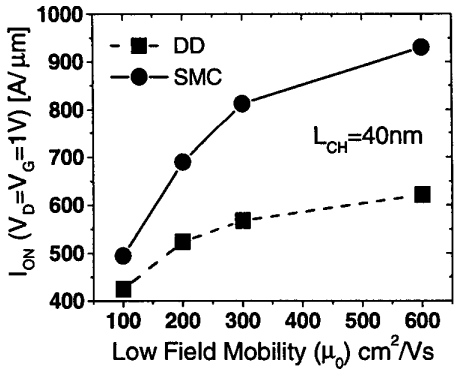

Fig. 6: I_{ON} as a function of μ_0 for a L_{CH} = 40nm nMOSFET.

4. pMOSFET Results

Fig. 7 shows pMOS I_{ON} simulated using MC and DD for V_G=V_D=1V and for two values of μ_0: 100 and 50 cm^2 / V-sec. These values span the value of hole inversion layer mobility for effective fields from 10^5 to 10^6 V/cm [11]; μ_0 = 50 is a typical value for a thin oxide nano-transistor. The differences between MC and DD are smaller here than in the nMOS case and only become noticeable for $L_{CH} \leq$ 80nm. For L_{CH} = 40nm, DD under-estimates I_{ON} by 25% for μ_0 = 100 and about 15% for μ_0 = 50. In contrast to the nearly ballistic nMOS case, the 40nm pMOS has a transmission probability of only 50% for μ_0 = 100 and about 40% for μ_0 = 50.

However, there is a strong velocity overshoot in the pMOS case towards the drain end (velocity exceeds 1.2e7 cm/s), and barrier lowering occurs. As in the nMOS case, the increased I_{ON} seen in the MC result is due to a combination of the increased density from barrier lowering and an increased velocity at the source.

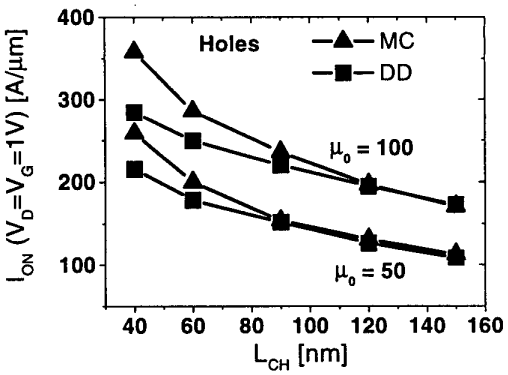

Fig. 7: pMOS I_{ON} as a function of L_{CH} for V_G=V_D=1V with μ_0 =100 and 50 cm^2 / Vs.

5. Optimizing Drift-Diffusion

For both the nMOSFET and pMOSFET the assumptions behind DD break down for L_{CH} < 100nm -- strong velocity overshoot near the drain for both devices and near ballistic transport in the nMOSFET lead to an underestimation of both I_{ON} and g_m. Surprisingly, the CT velocity expression is able to come close to the true velocity near the source, and so underestimates of I_{ON} are not as large as would be expected. This coincidence suggests that the agreement may be improved by adjusting the parameters β and v_{sat}. For nMOS with μ_0=300 the channel velocities can be fit over the entire L_{CH} range studied here using β=1.0 and v_{sat} = 2.2 10^7 cm/s (fig. 8). β must be reduced while v_{sat} is increased to maintain the correct velocity at low fields while capturing the overshoot at high fields.

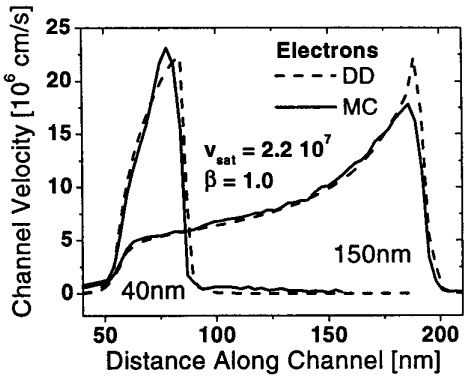

Fig. 8: Channel velocity for electrons (see fig. 2) with β =1, v_{sat}=2.2 10^7 cm/s in the DD model.

For holes the best fit is found for v_{sat} = 1.5e7 cm/s $\beta = 0.65$ for μ_0=100 and $\beta = 0.7$ for μ_0=50. I_{ON} corresponding to these values is shown in fig. 9. These values work surprisingly well for the mobilities and biases simulated here. It is unclear whether they are applicable in different MOSFET structures and under different operating conditions. It is possible that these parameters only work well when applied to the form of the potential found in MOSFETs since it is clear that these parameters yield unphysical velocities for high uniform electric fields in bulk.

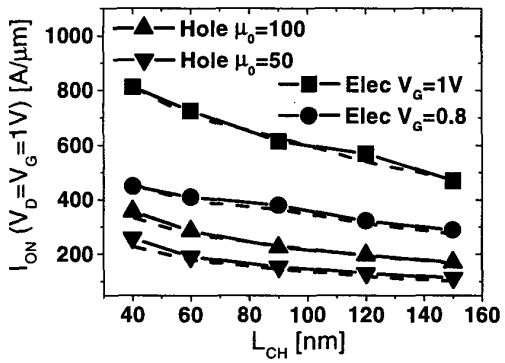

Fig. 9: I_{ON} for nMOS and pMOS with V_D=1V. Dashed lines: DD using β=1, v_{sat}=2.2 10^7 cm/s for electrons, β=0.65, v_{sat}=1.5 10^7 cm/s for holes; solid lines are MC. For electrons, μ_0=300 cm^2 /Vs; for holes V_G=1V.

6. Conclusions

In conclusion, drift-diffusion simulations using typical field-dependent mobility models (Caughey-Thomas) predict different scaling trends and discrepancies as large as 40% in nMOS I_{on}, 25% in pMOS I_{ON} and 50% in nMOS g_m. Monte Carlo simulation captures two non-local effects which explain the errors in drift-diffusion: higher near-ballistic injection velocities near the source and source barrier lowering which results from velocity overshoot in the channel. The errors are largest for electrons which have lower transport masses and lower scattering and are hence more ballistic. These errors may be partially compensated for by modifying the drift-diffusion field-dependent mobility model.

[1] S.E. Laux, M.V. Fischetti, IEDM Tech. Dig. p.877, 1997.
[2] T. Ghani et al, IEDM Tech. Dig. p.415, 1999.
[3] M. Mehrotra et al, IEDM Tech. Dig. p.419, 1999.
[4] G. Timp et al, IEDM Tech. Dig. p.55, 1999.
[5] See device model bench-marking in "Well-tempered Bulk Si-MOSFET," http://www-mtl.mit.edu/Well/.
[6] D.M. Caughey and R.E. Thomas, Proc. of the IEEE, p. 2192, Dec. 1967.
[7] J.A. Cooper, Jr. and D.F. Nelson, J. Appl. Phys., p.1445, March 1983.
[8] M.V. Fischetti, IEEE Trans. Elec. Dev. 38, p.634, 1991.
[9] R.W. Cohen and R.S. Muller, Solid State Elec., 23, p.35, 1978.
[10] J. Bude et al, IEDM Tech. Dig., 375, 1996 and refs. therein..
[11] E. Sangiorgi and M. Pinto, IEEE Trans. Elec. Dev., 39, p.356, 1992.
[12] S. Takagi et al, IEEE Trans. Elec. Dev. 41, p. 2357, 1994.
[13] M. Lundstrom, IEDM Tech. Dig. p.387, 1996.
[14] T. Kobayahi and K. Saito, IEEE Trans. Elec. Dev. 32, p.788, 1985