Extended Abstracts of the 1995 International Conference on Solid State Devices and Materials, Osaka, 1995, pp. 228-230

Invited

Monte Carlo Simulations of Impact Ionization Feedback in Sub-Micron MOSFET Technologies

Jeff D. Bude

AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974, USA

Monte Carlo transport simulation is used to clarify the nature of carrier heating in sub-micron MOSFET technologies, emphasizing processes leading to the population of the high energy tail of the distribution function. In particular, the important role of carrier heating by impact ionization feedback is demonstrated.

1. Introduction

Monte Carlo (MC) transport simulation is a widely recognized tool for detailed understanding of non-stationary transport in semiconductor devices. In particular, it is the only accurate method for obtaining the high energy tail of the electron energy distribution function (DF), responsible for impact ionization (II), oxide degradation and gate currents, I_G, in MOSFETs. Recently, MC MOSFET simulations have demonstrated that II feedback, the coupled II of electrons and holes, can have a strong effect on DF tails [1]. Although II feedback is recognized as an important current multiplication mechanism, its importance as a carrier heating mechanism has been largely overlooked. This work emphasizes the inclusion of II feedback in MC device simulations, and its implications for carrier heating in sub-micron MOSFET technologies.

2. Requirements for MC MOSFET simulation

MC simulation for the high energy tail of the DF requires the inclusion of several key features. First, it is well established that a full band structure representation is an essential starting point both because of its complex energy dispersion at high energies (even at low energies for holes) and the importance of a realistic density of states (DOS) for phonon scattering rates [2]-[3].

Another fundamental requirement is a realistic II scattering rate, $R_{II}(E)$, as a function of energy. MC simulators are typically calibrated first for bulk drift velocities, v_d, and II coefficients versus electric field. This does not uniquely determine either phonon or II scattering rates, and more importantly, the DF (eg., [2]). Two MC codes with different II scattering rates can both give the same bulk II coefficients and drift velocities with small changes to the phonon scattering rates; however, the DF functions in the two cases will be quite different (see fig. 1). Furthermore, in the highly non-stationary transport conditions generated in deep sub-micron MOSFETs, the II generated substrate current, I_B, is controlled by the II scattering rate, independent of the phonon rate; therefore, an accurate II rate is needed to model I_B [4]. This is discussed below.

Finally, II feedback must be included, especially for deep sub-micron MOSFETs. The secondary electrons and holes generated by II feedback can dominate DF tails, a fact that has recently been predicted and observed experimentally in deep sub-micron MOSFET technologies [1].

3. Full band MC MOSFET simulation

In the following, nMOSFETs of a 0.1 μm process [5] are simulated and compared with measurement. The devices are typical of deep sub-micron processes, having approximately 40 nm deep source/drain junctions and 5 nm thick gate oxides.

Simulations have been performed using the full-band MC transport simulator, SMC [6], as a post-processor for the device simulator PADRE [7]; electric fields are computed for a given bias condition by PADRE, and then, given these fields, SMC solves for the device DFs in a manner similar to [8]. SMC employs four pseudo-potential (pseudo-potential) conduction bands and three pseudo-potential valence bands stored on a tetrahedral mesh. The potential energy is stored on triangles, which with the band structure, forms the simplex discretization of phase-space used in SMC (Simplex MC), providing efficient and accurate numerics.

Optical and acoustic phonon scattering rates have been chosen to best approximate bulk drift velocity and II coefficients. They roughly reduce to the model in [9] for low energies and are extended consistently with the band structure to higher energies. Electron-electron (EE) scattering rates are treated using Thomas-Fermi screening with an effective electron temperature as discussed in [3]. The II scattering rates for electrons and holes, along with their final state secondary energy distributions have been computed from the pseudo-potential band structure of Si (see [10]).

WKB tunneling and MC transport [11] through the oxide are included for I_G calculation. Multiple rare-state algorithms are used to fully resolve the high energy DF tail. II feedback is treated using the iterative algorithm of ref. [12] which can greatly improve statistics for secondary DF tails. These statistical enhancement techniques are necessary, especially in computing I_G or device degradation, processes which require spatial resolution of DFs whose tails may decay by up to ten orders of magnitude before reaching the oxide barrier.

4. MOSFET channel heating

In the channel, there is a small amount of heating prior to reaching the pinch-off region (the smaller the channel length, L_{CH}, the more heating here), but most of the heating occurs in the pinch-off region where electric fields are high. MC simulations of these devices have shown that there is little thermalization of the channel electrons by EE scattering. As a consequence, channel carriers can gain at most an amount of energy limited to the channel potential energy drop.

Due to the shallow junctions and thin oxides used in the devices of the 0.1 μm process described above, the pinch-off electric fields are high enough (up to 1MV/cm for small V_{DS}) that transport through it is quasi-ballistic; carriers are injected into the drain with energies near V_p, the high field potential energy drop in the pinch-off region. In the drain they relax with nearly thermal tails. Fig. 2 shows the DF, F(E) and R_{II} as a function of energy. F(E) represents the total number density of carriers with a given energy in the simulation domain. As V_{DS} increases, V_p increases in parallel, and F(E) moves out rigidly with this potential increase. Sweeping V_{DS} samples R_{II} as a function of energy, a result directly measurable as substrate current, since $I_B = \int F(E)R_{II}(E)dE$. Fig. 3 shows the agreement between simulated and measured $I_{BR}=I_B/I_S$ versus V_{DS}. The clear agreement validates the II rate model, and the underlying channel DF shapes.

5. Heating by II feedback in MOSFETs

It was shown above that DFs due to channel transport are limited to energies below $V_p(V_{DS},V_{GS})$, suggesting that the high energy phenomena of oxide injection and trap formation may be suppressed by reducing supply voltages, V_{DS} in particular. However, the DF tail for energies greater than V_p can be populated by II feedback.

Fig. 4 illustrates the general phenomenon of II feedback in an nMOSFET. Channel electrons, e_1, are injected into the drain where they II forming low energy electron-hole pairs with current II multiplication M_1. The secondary electrons, e_2 leave through the drain while the secondary holes, h_2, diffuse to the drain-substrate junction (DBJ), are heated by its fields and are injected into the substrate where they II again with multiplication M_2 forming e_3 and h_3. The h_3 holes leave through the substrate, but the e_3 electrons fall back through the DBJ and vertical gate controlled potential drops reaching the oxide interface. This process continues with e_3 ionizing leading to a series of pair productions alternating between electrons and holes (II feedback) with multiplications M_3, M_4, etc. Here, $I_B= I_S(1+ M_1+ M_1M_2+ M_1M_2M_3+...)$. The DBJ is in breakdown when this series diverges, but in the following, the devices are not in breakdown and all the $M_i<1$.

A L_{CH}=0.5μm device of the 0.1 μm process described above has been simulated to quantify the feedback effect. Fig. 5 shows the potential energy along the channel (from A to B in fig. 4) and from the drain to the substrate (B to C). Here, $V_{GS}=V_{DS}=3V$. The channel transport has been described above, and the DF integrated over the oxide interface is shown in fig. 6. The e_1 DF shows a rapid decay for energies past V_p = 2.1 eV, so there are no channel electrons above the oxide conduction band discontinuity of about Δ_{ox}=3.1eV, but there are many at the II rate threshold of about 1.1eV, so $I_B\neq0$.

The simulated e_3 DF shown in fig. 6 demonstrates that the e_3 DF extends to much higher energies than the 2.1eV limit of the e_1 DF, even though the current carried by them is much less ($I_{e_3}/I_{e_1}=M_1M_2$). In fact, there are a large number of e_3 electrons with energies above Δ_{ox}. The DF shapes of higher order secondaries such as e_5 are fairly close to that of the e_3 DFs, but their magnitudes are much less if the DBJ is not in breakdown. Therefore, it is sufficient to consider only the e_3 secondary electron DF.

The fact that the feedback generated e_3 DF dominates the channel e_1 DF at high energies is a general result. The energies of e_3 electrons reaching the oxide interface can be as great as $E_3^{max} =qV_{DB}+qV_{bi}+E_{sec}$, where E_{sec} is the energy of formation of the e_3 electrons by the II of the h_2 holes, and qV_{bi} is the built-in junction potential. Typically, $qV_{bi}\approx$ 0.5eV to 0.8eV, and $E_{sec}>0$. (It is important to have an accurate model for E_{sec} as it can have a large effect on the DF tail, especially for low qV_{DB}.) Clearly, $E_3^{max} >V_{DS}>V_p$. Hence, the e_3 DF determines the total electron DF for energies greater than $\approx V_p$, and possibly for lower energies, depending on the relative magnitudes of the channel and DBJ fields.

The resulting DF is bi-modal; the channel DF is quasi-ballistic decaying rapidly above V_p, and the feedback DF extends to high energies. For the conditions of fig. 5, the e_1 DF dominates below 2V controlling the II generated I_B which has a threshold near 1.1eV, but the e_3 DF controls the tail and is responsible for I_G and oxide trap formation.

6. Gate currents by II feedback

It is clear from above that V_{BS} should have a large effect on the tail of the DF. When a negative bias of V_{BS} is applied, changes in the channel potential are small, but qV_{DB} changes rigidly with V_{BS} (see fig. 5). Therefore, e_1 heating and I_B are not strongly affected, whereas e_3 heating and I_G should change exponentially. The DFs for the devices of fig. 5 with V_{BS}=−1.5 are also shown in fig. 6. This is confirmed experimentally in fig. 7 which shows $I_{GR}=I_G/I_S$ versus $I_{BR}=I_B/I_S$ for $V_{DS}=V_{GS}$ at various V_{BS}. At a given V_{DS} (eg., dashed line: V_{DS}=3.5V), I_G increases exponentially while I_B increases only linearly. Also, for a given value of I_{BR}, I_{GR} is a strong function of V_{BS}. This result contradicts the "effective temperature models" which predict a direct correlation between I_{GR} and I_{BR} [13] based on the assumption that both currents are generated by a thermalized channel DF.

Finally, fig. 8 shows a comparison between simulated and measured I_{GR} and I_{BR} for various bias configurations. Qualitative trends are clearly reproduced by the simulation. Due to the admittedly approximate treatment of the difficult topic of oxide injection, quantitative agreement with experiment is more questionable.

7. Conclusions

Physically based MC transport simulation has clarified the formation of high energy DF tails in sub-micron MOSFETs, showing good agreement with experimental I_B and I_G measurements. In particular, the importance of a full-band structure, a realistic II rate and the inclusion of II feedback have been demonstrated.

8. References

[1] J.D. Bude, 1995 Symposium on VLSI Technology (1995) 125.
[2] J.Y. Tang, K. Hess, J. Appl. Phys. **54** (1983) 5139.
[3] M.V. Fischetti, S.E. Laux Phys. Rev. B **38** (1988) 9721.
[4] J.D. Bude, M. Mastrapasqua, submitted to IEEE EDL (1995).
[5] K.F. Lee, et al., IEDM Tech. Digest (1993) 131.
[6] J. Bude, R.K. Smith, in Semiconductor Science and Technology, HCIS 8 **9 5S** (1994) 840.
[7] M.R. Pinto, et al., IEDM Tech. Digest (1992) 923.
[8] J.M. Higman et al., IEEE Trans. Elec. Dev. **ED-36** (1989) 930.
[9] C. Canali et al, Phys. Rev. B **12** (1975) 2265.
[10] E.O. Kane, Phys. Rev. **159** (1967) 624.
[11] M.V.Fischetti, et al. Phys. Rev. B (1985) 8124.
[12] I. C. Kizilyalli, J. Bude, IEEE Trans. Elec. Dev. **41** (1994) 1083.
[13] S. Tam, et al., IEEE Trans. Elec. Dev. **31** (1984) 1116.

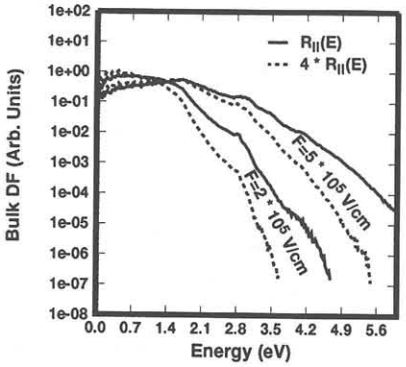

Figure 1: Bulk DFs using different II rates, R_{II} from fig.2 and 4*R_{II}, with slightly different phonon scattering rates. Bulk II coefficients are identical for

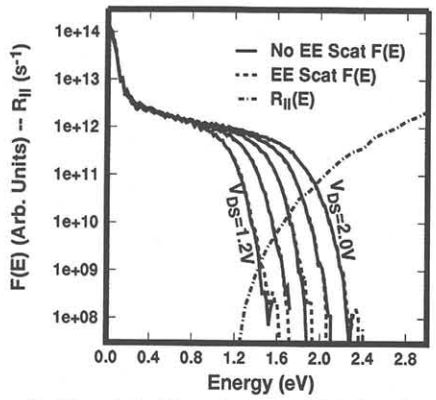

Figure 2: Simulated hot electron DFs in a L_{CH}=1μm device of a 0.1μm process and II rate versus energy.

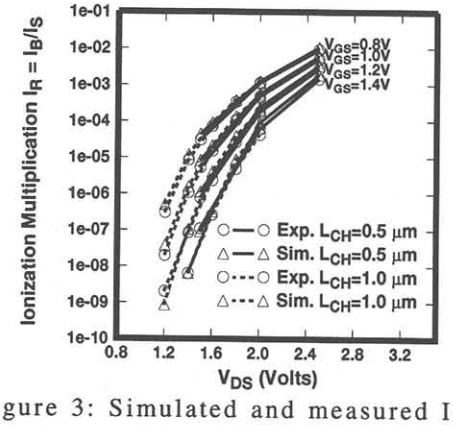

Figure 3: Simulated and measured I_{BR} for L_{CH}=1μm and 0.5μm devices of a 0.1μm process.

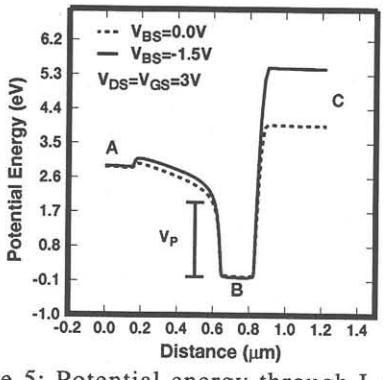

Figure 5: Potential energy through L_{CH}=0.5μm device of a 0.1μm process from A-B, B-C (fig. 4).

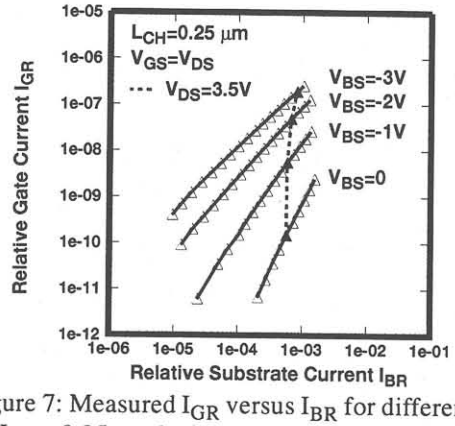

Figure 7: Measured I_{GR} versus I_{BR} for different V_{BS} on L_{CH}=0.25μm device.

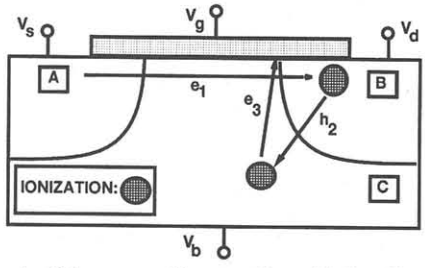

Figure 4: Diagram illustrating II feedback in MOSFETs.

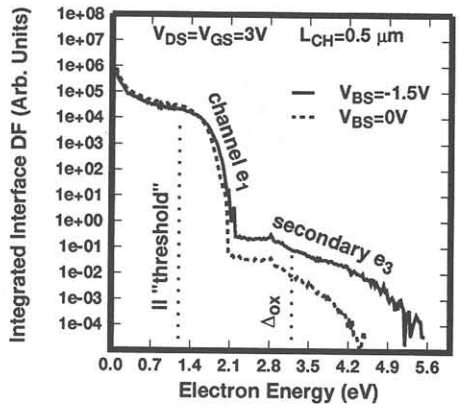

Figure 6: DFs integrated at the oxide interface for conditions of figure 5.

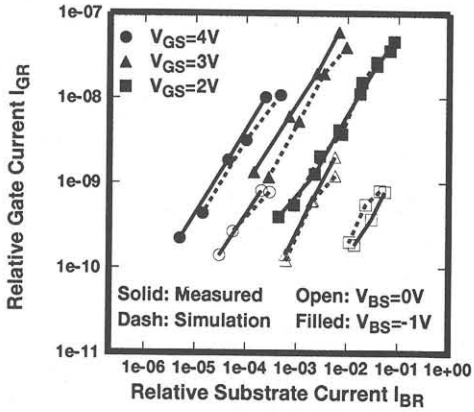

Figure 8: Measured and simulated I_{GR} and I_{BR}. Open symbols, V_{BS}=0V; filled symbols V_{BS}=-1V.