Essential Physics of Carrier Transport in Nanoscale MOSFETs

Mark Lundstrom, *Fellow, IEEE*, and Zhibin Ren

***Abstract*—The device physics of nanoscale MOSFETs is explored by numerical simulations of a model transistor. The physics of charge control, source velocity saturation due to thermal injection, and scattering in ultrasmall devices are examined. The results show that the essential physics of nanoscale MOSFETs can be understood in terms of a conceptually simple scattering model.**

***Index Terms*—Charge carrier processes, MOSFETs, semiconductor device modeling, semiconductor devices, transistors.**

I. Introduction

SCALING MOSFETs to their limits is a key challenge now faced by the semiconductor industry. Physically detailed simulations which capture the off-equilibrium transport (e.g., velocity overshoot) [1]–[3] and the quantum mechanical effects that occur in these devices [4] can complement experimental work in addressing these challenges. Also needed, however, is a simple conceptual view of the nanoscale transistor—to help interpret detailed simulations and experiments and to guide experimental work. Such a model has recently been outlined [5], [6]. Our objective in this paper is to critically examine this simple view through the use of two-dimensional (2-D) numerical simulations. As a vehicle for these studies, we use a model 10-nm double-gate (DG) MOSFET, but we expect the conclusions to apply to nanoscale MOSFETs more generally. We use a semiclassical approach, because recent work shows that MOSFETs operate essentially classically down to channel lengths of about 10 nm [7], [8]. We also restrict our attention to the steady-state current–voltage (I–V) characteristics, which are relevant to the high-speed operation of digital circuits [9].

Fig. 1 summarizes the essential physical picture that will be discussed in this paper. We adopt a transmission view of the device [6], [10] in which carriers are injected into the channel from a thermal equilibrium reservoir (the source), across a potential energy barrier whose height is modulated by the gate voltage, into the channel, which is defined to begin at the top of the barrier. The beginning of the channel is populated by carriers injected from the thermal equilibrium source (and, under low drain bias, from the thermal equilibrium drain). The density of carriers at the top of the barrier is controlled by MOS electrostatics so that the charge in the semiconductor balances that in the gate. Under equilibrium conditions ($V_{DS} = 0$ V) in an electrostatically well-tempered device, equilibrium one-dimensional (1-D) MOS electrostatics apply at this point, so the inversion layer density can be computed as for a 1-D MOS capacitor. Above threshold, therefore

$$Q_i(0) = qn_s(0) \approx C_{\text{eff}}(V_{GS} - V_T) \tag{1}$$

where C_{eff} is the effective oxide capacitance (as influenced by quantum mechanical confinement, polysilicon depletion, finite density-of-states, etc. [9]). We will show that a type of "gradual channel approximation" applies at this point, so that the inversion layer density at the source end of the channel remains nearly equal to its equilibrium value even when a drain bias is applied.

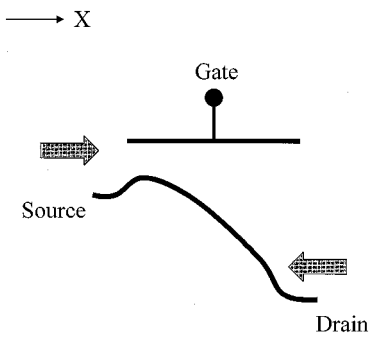


Fig. 1. Conduction subband edge versus position from the source to the drain of a nanoscale MOSFET under high gate and drain bias. Also shown are the thermal injection fluxes from the source and drain.

Some fraction of the carriers injected from the source into the channel backscatter and return to the source; others flow out the drain and comprise the steady-state drain current I_D. (For a high drain bias, carriers injected from the drain need not be considered). Assuming current continuity, I_D may be evaluated at the beginning of the channel where the carrier density is known from MOS electrostatics to find

$$I_D = WQ_i(0)\langle v(0)\rangle \approx WC_{\text{eff}}(V_{GS} - V_T)\langle v(0)\rangle \tag{2}$$

where $\langle v(0)\rangle$ is the average velocity of carriers at the beginning of the channel. The maximum value of $\langle v(0)\rangle$ is approximately the equilibrium uni-directional thermal velocity $\tilde{v}_T$, because the positive velocity carriers at the beginning of the channel were injected from the thermal equilibrium source [5]. Backscattering from the channel determines how close to this upper limit the device operates. Under high drain bias, the average velocity at the beginning of the channel can be related to a channel backscattering coefficient r, according to [5]

$$\langle v(0)\rangle \approx \left(\frac{1-r}{1+r}\right)\tilde{v}_T \tag{3}$$

Manuscript received May 15, 2001; revised September 25, 2001. This work was supported by the Semiconductor Research Corporation and by the National Science Foundation, Distributed Center for Advanced Electronics Simulations. This paper is an expanded version of a recent conference presentation [32]. The review of this paper was arranged by Editor M. J. Deen.

The authors are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: lundstro@purdue.edu).

Publisher Item Identifier S 0018-9383(02)00227-7.

where $0 < r < 1$ is a backscattering coefficient in the spirit of McKelvey [11], [12]. (Note that when (3) is inserted into (2), we get a result presented earlier [5]. Also note that the backscattering coefficient r depends on the scattering physics and on the self-consistent potential within the channel, so r is a function of the gate and drain biases). The importance of the source velocity is, of course, well known (e.g., [13]); we relate it to a channel backscattering coefficient in order to clarify the source velocity limit.

Because of the high electric field and strong velocity overshoot, carrier transport through the drain end of the channel is rapid. As a result, the dc current is controlled by how rapidly carriers are transported across a short low-field region near the beginning of the channel. Carriers diffuse across the beginning of the channel in much the same way that they diffuse across the base of a bipolar transistor, and they are collected by the high-field portion of the channel much as in the collector of a bipolar transistor [14]. We refer to the critical, low-field region near the beginning of the channel as the "kT-layer" because it is roughly the distance over which the channel potential drops by k_BT/q. Scattering within the kT-layer limits the steady-state drain current; scattering near the drain end of the channel has only an indirect effect. This is analogous to the well-known Bethe condition for thermionic emission in a forward-biased metal-semiconductor diode [15], although the physics is different because in a MOSFET, the flow of carriers is down the potential barrier rather than up. For well-designed MOSFETs, the length of the kT-layer (which is set by 2-D electrostatics as influenced by velocity overshoot within the channel [16]) is about one-mean-free path, which means that transport across this layer is quasi-ballistic.

In the following sections, we use detailed, numerical simulations to confirm this basic physical picture and to expand upon it. Note that in presenting the basic physical picture, we have made several simplifying assumptions. For example, we assumed high drain bias, although a full range expression can be developed [17]. We also assumed nondegenerate carrier statistics; degeneracy increases the average thermal velocity, causes the average velocities of the positive and negative halves of the distribution at the top of the barrier to differ, and influences the length of the critical region (i.e., the criterion of a k_BT/q potential drop must be generalized for degenerate statistics). Some of these issues will be discussed further in this paper, but our intent is to present the basic physical picture in simple form, so a full discussion must be deferred to later publications. The following specific issues will be addressed in this paper:

1) injection velocity limits at the source end of the channel;
2) the off-equilibrium distribution function at the source;
3) charge control in a nanoscale MOSFET;
4) the role of scattering and the generalized Bethe condition for a MOSFET;
5) the role of velocity overshoot in the channel;
6) the magnitude of the quantum contact resistance in nanoscale MOSFETs.

To examine these effects, we numerically simulated the simple model MOSFET shown in Fig. 2. The device is a DG MOSFET with an exceptionally thin (1.5 nm) Si body, a 1.5-nm SiO_2 gate oxide, and $L_G = 10$ nm. A hypothetical midgap work function gate material was assumed. The device is assumed to be

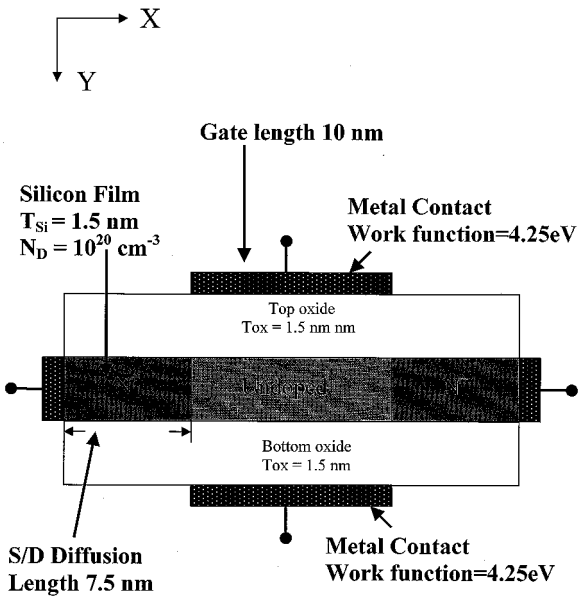


Fig. 2. Structure of the $L = 10$ nm DG MOSFET with $t_{ox} = 1.5$ nm, $t_{\rm Si} = 1.5$ nm, and $V_DD = 0.6$ V. This device was simulated with a 2-D solution to Poisson's equation coupled to a 1-D transport solution [8].

wide in the z-direction (out of the page), so that many transverse modes are occupied. Also note that the idealized metal contacts in Fig. 2 approximate the actual contacts where dissipative scattering would dominate and maintain a thermal equilibrium carrier distribution. (Real contacts would also flare out to reduce series resistance).

Our simulations treated electrostatics two-dimensionally, but transport is essentially 1-D in this geometry, so a simplified 1-D transport model was used [8]. Quantum confinement effects in the direction normal to the Si film were treated in the one subband approximation. Several different approaches were used to describe transport along the channel. In the ballistic case, both a semiclassical (Boltzmann) solution and a quantum solution using a Green's function approach [18] were used. Quantum transport in the presence of phase breaking scattering was treated using a simple generalization of the Büttiker probe concept [18] (we verified that this approach captured the essential features of scattering observed in semiclassical approaches). Conventional drift-diffusion and energy transport models were also available.

The simplified device geometry and the ultrathin body help to clarify the device physics to be explored in this study, but the conclusions of this study are born out by full 2-D simulations of thicker body devices. Those results, however, are clouded by multi-subband conduction and stronger 2-D electrostatics (e.g., DIBL). Although the model device is a DG MOSFET, we expect that the general conclusions of the study will apply to bulk MOSFETs as well. Fig. 3 shows the computed self-consistent conduction subband profiles versus position under a variety of bias conditions. It is important to note that we plot the bottom of the first subband versus position, not the conduction band edge, and that when we speak of the "source-to-channel barrier," we refer to the subband energy versus position—not to the conduction band profile. (The program used to perform these simulations is available in [19], and more extensive simulations of the same device have been reported in [8]).

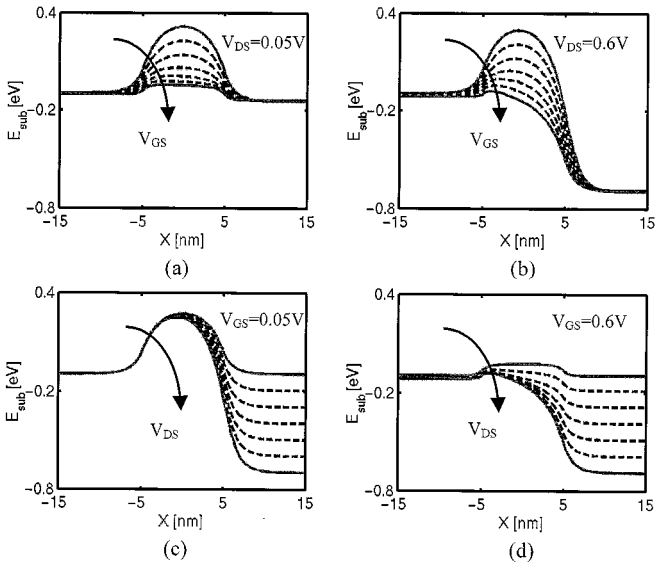


Fig. 3. Computed self-consistent conduction subband edge versus position for DG MOSFET of Fig. 2: (a) $V_{DS} = 0.05$ V and V_{GS} from 0.0 V to 0.6 V; (b) $V_{DS} = 0.6$ V and V_{GS} from 0.0 V to 0.6 V; (c) $V_{GS} = 0.05$ V and V_{DS} from 0.0 V to 0.6 V; and (d) $V_{GS} = 0.6$ V and V_{DS} from 0.0 V to 0.6 V.

II. The Ballistic MOSFET

The physical picture presented in Section I is most easily examined in the ballistic limit, and since present-day devices operate relatively close to this limit [20], [21], there is also a practical motivation to examine the ballistic MOSFET. The ballistic MOSFET was first examined by Natori, who developed an analytical model [22]. For this paper, we numerically simulated the model MOSFET of Fig. 2 using a semiclassical, ballistic transport model coupled to a 2-D solution to Poisson's equation [23].

A. Source Velocity Limits in a Ballistic Nano-MOSFET

In Section I, we argued that the maximum average carrier velocity at the beginning of the channel was the equilibrium, uni-directional thermal velocity, as first pointed out by Natori [22]. Assuming that only one subband is occupied, it can be shown that [22], [24]

$$\tilde{v}_T = \sqrt{\frac{2k_BT_L}{\pi m_t^*}}\left\{\frac{\mathcal{F}_{1/2}(\eta)}{\ell n(1+e^{\eta})}\right\} = v_T\left\{\frac{\mathcal{F}_{1/2}(\eta)}{\ell n(1+e^{\eta})}\right\} \quad (4)$$

where $\eta = (E_F - \varepsilon_1)/k_BT$, and the factor in brackets accounts for carrier degeneracy and approaches unity for a nondegenerate gas. (More generally, when multiple subbands are occupied, Schrödinger–Poisson simulations are needed [24]). Fig. 4 shows the equilibrium $\tilde{v}_T$ versus n_S characteristic computed from (4). Note that below threshold, $\tilde{v}_T \approx v_T \approx 1.2 \times 10^7$ cm/s, but that above threshold, the carriers become degenerate, and the thermal injection velocity increases. Finally, note that the degenerate thermal injection velocity is the average velocity of all the carriers, while the Fermi velocity v_F refers to the velocity of carriers at the Fermi level. The two are related by

$$\tilde{v}_T(\eta \to \infty) = \left(\frac{4}{3\pi}\right) v_F. \quad (5)$$

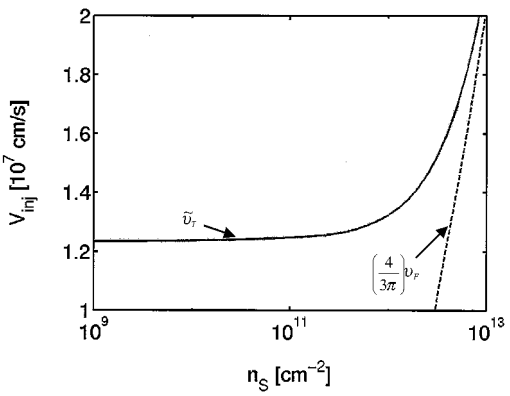


Fig. 4. Euilibrium thermal injection velocity $\tilde{v}_T$ versus inversion layer density n_S for the DG SOI MOSFET as evaluated from (4). Also shown is v_F, the Fermi velocity.

We assert that the equilibrium, uni-directional thermal velocity is the maximum velocity that can be observed at the source end of the channel. The maximum source velocity exceeds the saturated velocity [23], but the origin of this high velocity is much different than that of the conventional velocity overshoot that occurs in steep electric field gradients [25]. These high-source velocities will, however, not be achieved unless the velocity within the channel is even higher (e.g., unless strong velocity overshoot within the channel).

The simulations displayed in Figs. 5 and 6 confirm the assertions made in the previous paragraph. Fig. 5 is a plot of $v_{\mathrm{inj}} = \langle v(0)\rangle$ versus drain bias as obtained by simulating the ballistic device of Fig. 2. (The location $x = 0$ is taken as the top of the source-to-channel barrier, which changes with bias). Under low bias, the average velocity is nearly zero because the negative velocities of carriers injected from the drain nearly cancel the positive velocities of those injected from the source. When the drain bias exceeds a few k_BT/q, then the negative velocity carriers injected from the drain are suppressed, and the average velocity saturates at the equilibrium thermal velocity $\tilde{v}_T$. Fig. 6 shows the average velocity versus position profiles at different drain to source voltages. As expected in this ballistic transistor, the velocity near the drain increases without limit (band structure limits have not been included). Under high drain voltages, however, the velocity at the top of the barrier saturates at the value displayed in Fig. 4. These results confirm the assumption made in Section I and earlier [5]. They show that velocity saturation occurs in a ballistic MOSFET, but it is the velocity at the top of the barrier that saturates at the thermal limit as opposed to the high-field velocity saturation in a bulk semiconductor which occurs because of scattering.

In the ballistic MOSFET, a special kind of equilibrium exists; **k**-states are in equilibrium with the contact from which they were populated [10]. The overall carrier distribution, however, can have a highly off-equilibrium shape. For example, under high drain bias, the carrier distribution at $x = 0$ assumes a hemi-Fermi–Dirac shape. This is suggested by the dashed line in Fig. 5, which shows the ratio J^-/J^+ of the negative flux to the positive flux versus drain bias. This ratio approaches zero when the drain bias is large enough to suppress the injection of negative-velocity carriers from the drain. The net velocity then

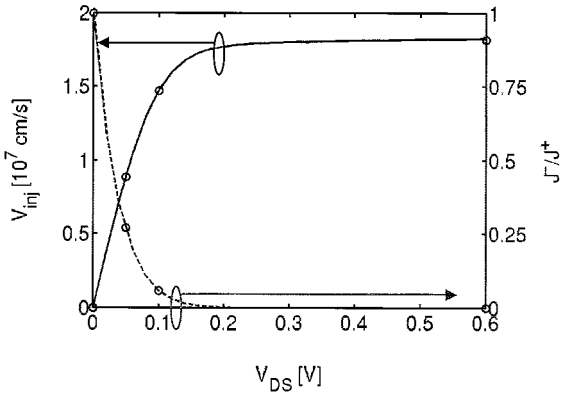


Fig. 5. Average velocity at the beginning of the channel versus V_{DS} for the device of Fig. 2 under ballistic conditions. For the gate voltage used $n_S \approx 5 \times 10^{12}$ cm^{-2}. Also shown is the ratio J^-/J^+ (negatively-directed flux to the positively-directed flux), which is a measure of the anisotropy of the distribution (dashed line). Note that the velocity at the beginning of the channel saturates at the thermal equilibrium injection velocity as given by (4) when the negative half of the distribution is suppressed $J^-/J^+ = 0$). The large dots identify the four voltages examined in Fig. 7.

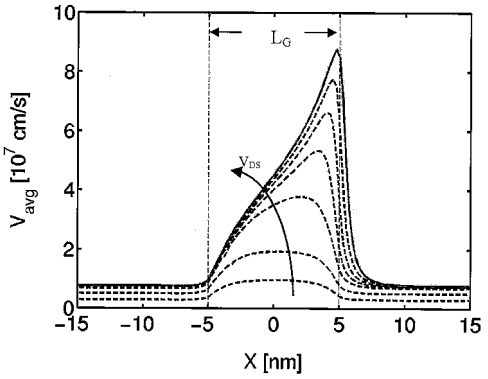


Fig. 6. Average velocity versus position for the device of Fig. 2 under ballistic conditions. For the gate voltage used $n_S \approx 5 \times 10^{12}$ cm^{-2}. Results for several different drain voltages are shown.

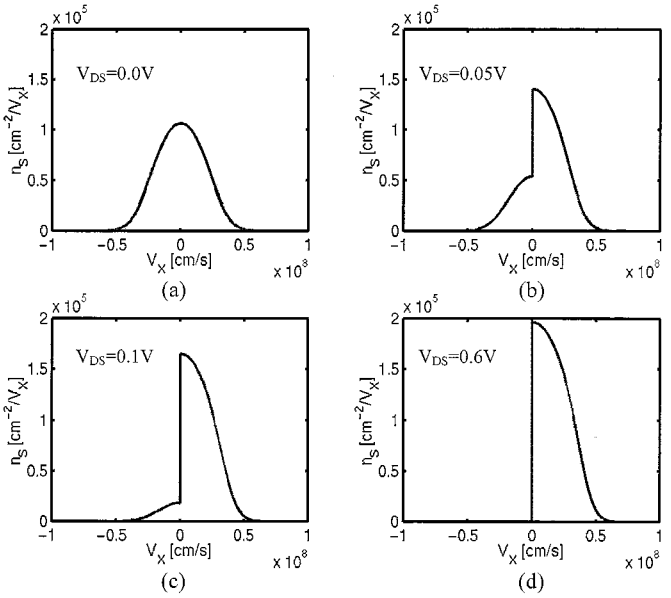


Fig. 7. Computed, ballistic velocity distributions at the top of the source to channel barrier under (a) $V_{DS} = 0.0$ V; (b) $V_{DS} = 0.05$ V; (c) $V_{DS} = 0.1$ V; and (d) $V_{DS} = 0.6$ V. For each case, $V_{GS} = 0.6$ V. The units are such that the area under the curves is the electron density.

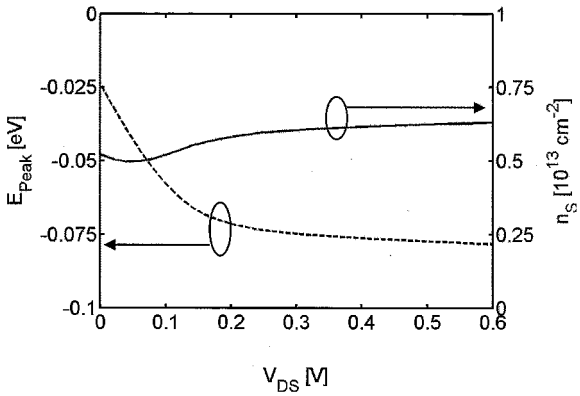


Fig. 8. Illustration of the charge control mechanism for the device of Fig. 2 under ballistic conditions. Solid line: the carrier density at the beginning of the channel versus V_{DS} for the device. Dashed line: the source to channel barrier height versus V_{DS}. Fig. 5 showed that as the ratio J^-/J^+ decreases from 1 to 0, the average velocity increased. $n_S(0)$ remains essentially constant and the source-to-channel barrier height decreases with increasing V_{DS} to maintain a constant carrier density at the top of the barrier. (The small increase can be attributed to DIBL.)

saturates at $\tilde{v}_T \approx 1.8 \times 10^7$ cm/s which is 5% higher than the equilibrium thermal injection velocity shown in Fig. 4 (the difference is due to 2-D electrostatics). An extensive study of distribution function effects in ballistic MOSFETs has recently been reported [23]. What happens at the top of the barrier is shown in Fig. 7, which plots the computed ballistic distribution functions at the top of the barrier for the four different voltages noted in Fig. 5. For low V_{DS}, the velocity distribution is nearly symmetrical about $v_x = 0$. (In a long-channel device, this symmetry is a result of carrier scattering, but in the ballistic MOSFET, the positively-directed carriers are injected from the source and the negatively-directed carriers from the drain). As the drain bias increases, the magnitude of the negative-velocity component decreases. Note, however, that although the overall velocity distribution has a highly nonequilibrium shape, each half is in equilibrium with its respective contact.

B. Charge Control in a Ballistic Nano-MOSFET

We turn now to the issue of charge control in the ballistic nanotransistor. Because the carrier distribution at the top of the barrier approaches a hemi-Fermi–Dirac distribution under high drain bias, it might be expected that under high bias, $n_S(0)$ would be one-half of its equilibrium value [see (1)]. Fig. 8, however, shows that this is not the case—$n_S(0)$ is approximately constant with drain bias. This occurs because MOS electrostatics demands that the charge on the gate balance that in the semiconductor, so that as V_{DS} increases, the conduction band is pushed down, more electrons are injected from the source, and $n_S(0)$ is maintained approximately at the value given by (1). [This barrier lowering is also seen in Fig. 3(d)]. The plot of $n_S(0)$ versus V_{DS} in Fig. 8 confirms that in a "well-tempered MOSFET," which is designed to electrostatically isolate the drain from the source [26], MOS electrostatics maintains the inversion layer charge at the beginning of the channel at an approximately constant value. Although the velocity distribution

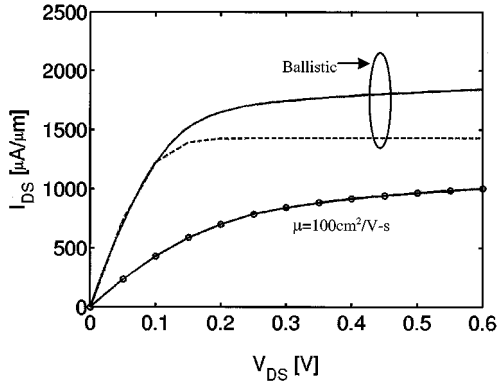


Fig. 9. Comparison of the simulated I_{DS}–V_{DS} characteristics of the ballistic device with the analytical model of (6). Solid line: simulated ballistic I_{DS} versus V_{DS} for a gate voltage of 0.6 V. Dashed line with symbols: analytical I_{DS} versus V_{DS}. Solid line with symbols: simulated I_{DS} versus V_{DS} including the effects of scattering. (An inversion layer mobility of 100 $\text{cm}^2/\text{V-s}$ was assumed.)

is highly nonequilibrium in shape, the charge density is maintained at approximately its equilibrium value. The same effect has also been observed in 2-D Monte Carlo simulations [27].

C. The Channel Resistance of a Ballistic Nano-MOSFET

Because the physics of the ballistic MOSFET is rather simple, a compact model is readily developed. Following Natori [22] (see also [17] and [24]) and assuming single subband occupation, one can show that

$$\frac{I_D}{W} = Q_i(V_{GS})\tilde{\upsilon}_T \left[\frac{\frac{1-\mathcal{F}_{1/2}(\eta-U_{DS})}{\mathcal{F}_{1/2}(\eta)}}{\frac{1+\ell n\left(1+e^{\eta-U_{DS}}\right)}{\ell n(1+e^{\eta})}} \right] \tag{6}$$

where $Q_i(V_{GS})$ is the inversion layer charge (approximately $2C_{\text{eff}}(V_{GS}-V_T)$ above threshold) and U_{DS} is V_{DS} normalized to k_BT/q. (Under nondegenerate conditions, the Fermi–Dirac integrals are replaced by exponentials, and under high drain bias, the term in brackets approaches unity). Under high gate bias $Q_i \approx 2C_{\text{eff}}f(V_{GS}-V_T)$, so (6) reverts to (2) with $\langle \upsilon(0) \rangle = \tilde{\upsilon}_T$..

Conventionally, a MOSFETs channel resistance is proportional to its channel length, but there is also a ballistic component independent of channel length that may be important in nanoscale MOSFETs [20]. For low drain bias, (6) gives the ballistic conductance as

$$\frac{G_{DS}}{W} = \frac{I_{DS}}{V_{DS}} = Q_i(V_{GS}) \left(\frac{\tilde{\upsilon}_T}{\frac{2k_BT}{q}} \right) \left[\frac{\mathcal{F}_{-1/2}(\eta)}{\mathcal{F}_{1/2}(\eta)} \right]. \tag{7}$$

As discussed in [24], under fully degenerate conditions, (7) reduces to $G_{DS} = M(e^2/2h)$, where M is the number of occupied transverse modes.

In Fig. 9, we compare the ballistic I–V characteristics as computed by direct numerical simulation and by the analytical expression (6). The agreement is good—except for the output conductance, a 2-D effect not treated by the 1-D analytical model. The channel resistance R_{DS} of this nano-MOSFET, as computed from the slope of the simulated characteristic in Fig. 9 or from (7), is about 60 Ω-μm. For comparison, we also show the simulated I_{DS} versus V_{DS} characteristic for the transistor including a simple model for scattering (to be discussed in Section III). With scattering included, the channel resistance increases to about 200 Ω-μm. This value includes the conventional channel resistance, which is proportional to channel length L and the quantum contact resistance, which is given by (7) and is independent of L. Note that the ballistic channel resistance is about 30% of the total channel resistance. Depending on the channel length and inversion layer mobility, this length-independent component to R_{DS} may become important.

III. Scattering

In a ballistic MOSFET, the positive-velocity carriers at the top of the barrier are injected from the source and negative-velocity carriers from the drain, but scattering mixes these two streams. The result is that the carrier distribution at the top of the barrier does not approach a hemi-Fermi–Dirac distribution under high drain bias; $\langle \upsilon(0) \rangle$ is less than $\tilde{\upsilon}_T$ under on-current conditions. When $V_{DS} \gg k_BT/q$, so that all negative-velocity carriers at the top of the barrier arise from backscattering, (3) applies. Well-designed MOSFETs currently operate with $r \approx 0.4$ [20], so from (3), $\langle \upsilon(0) \rangle$ is about one-third of its limit, but devices with $r \approx 0.2$ have been reported [21]. Fig. 9 illustrates how scattering reduces device performance with respect to the ballistic limit; the channel resistance increases to several times the ballistic resistance, the on-current is reduced to about one-half of the ballistic limit, the drain saturation voltage increases, and the output conductance increases. (Note that we only treat scattering within the channel so that well-understood series resistance effects do not cloud the results).

In this section, we examine two issues in detail:

1) charge control in the presence of scattering; and
2) why the channel backscattering coefficient is sensitive to backscattering near the source end of the channel and relatively insensitive to scattering deep within the channel.

For these studies, we use a Green's function method with a simple, Büttiker-probe model of scattering, which we tested to ensure that it captures the essential physics of scattering in a MOSFET. As shown in [8], device operation is essentially classical (except for the strong quantum confinement effects); the quantum transport model was used because it was available and had been extensively tested on this device [8]. The broadening parameter η in the scattering model [18] was set to 30 meV, which results in an inversion layer mobility of 100 $\text{cm}^2/\text{V-s}$ for a long-channel device. For a discussion of the formalism and solution methods, the reader is referred to [18].

A. Charge Control and Velocity Saturation in the Presence of Scattering

Fig. 10, which compares the self-consistent conduction subband profiles under on-current conditions with and without scattering, shows that the source-to-channel barrier is higher in the presence of scattering. This can be understood in terms of the self-consistent electrostatics of the MOSFET. For a given gate voltage, we expect the same inversion layer charge density at the top of the barrier—in the presence or absence of scattering. For

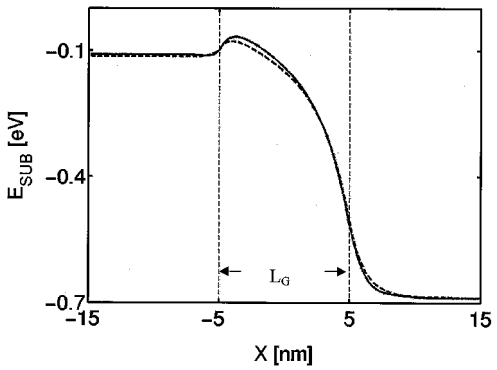


Fig. 10. Illustration of the effects of scattering on the self-consistent potential within the device of Fig. 2 under a bias of $V_{DS} = V_{GS} = 0.6$ V. Solid line: the lowest conduction subband energy versus position in the presence of scattering. Dashed line: the same plot in the absence of scattering. The key difference is a slightly lower source-to-channel energy barrier in the presence of scattering.

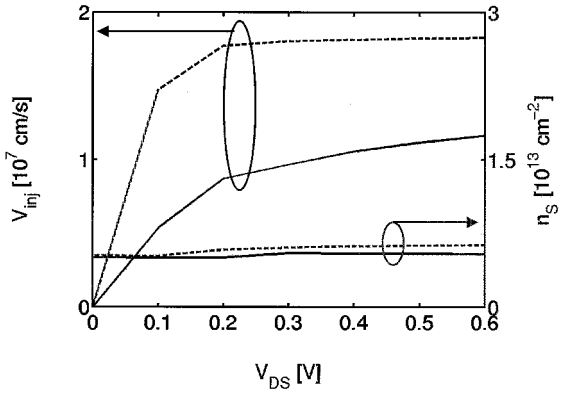


Fig. 11. Illustration of the effects of scattering on the average velocity and charge at the top of the barrier for the device of Fig. 2 with $V_{GS} = 0.6$ V. The carrier density at the beginning of the channel versus V_{DS} (right vertical axis). The average velocity at the beginning of the channel versus V_{DS} (left vertical axis). The solid lines include scattering, and the dashed lines are the corresponding results for ballistic conditions (from Figs. 5 and 8).

the ballistic case, the carrier distribution is a hemi-Fermi–Dirac distribution, and the barrier height is established to provide the necessary inversion layer density. In the presence of scattering, the carrier distribution function at the top of the barrier is more closely symmetric in v_x; so a higher barrier results in the same inversion layer density.

Fig. 11 displays the simulated average velocity and carrier density at the top of the barrier versus V_{DS} with a high gate voltage applied. The corresponding results for the ballistic case (from Figs. 5 and 8) are also displayed. First of all, note that the inversion layer density at the top of the barrier is nearly equal to its equilibrium value in the presence or absence of scattering (this is a simple consequence of self-consistent MOS electrostatics and is relatively insensitive to the specific transport model). Note also that the maximum velocity at the top of the barrier does not saturate as clearly as for the ballistic case and that it is well below the thermal injection limit. Still, one can identify a drain saturation voltage of $V_{DSAT} \approx 0.3$ V, which is greater than the $\approx$0.2 V in the ballistic case. It is clear that the mechanism for velocity saturation at the top of the barrier is different in the case of scattering and that it does not involve suppression of carrier injection from the drain as in the ballistic case.

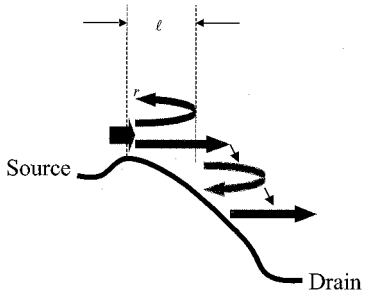


Fig. 12. Illustration of carrier backscattering in a MOSFET under high drain bias. If a carrier backscatters beyond the critical distance ℓ from the beginning of the channel, then it is likely to exit from the drain and unlikely to return to the source.

In the presence of scattering, velocity saturation at the beginning of the channel occurs because of the self-consistent electrostatics in the device. As shown in Fig. 3(d), for V_{DS} greater than about 0.3 V, most of the additional applied drain voltage is dropped across the drain end of the channel, and conditions near the source are relatively constant. From (3), one can estimate that $r \approx 0.3$ at $V_{DS} \approx V_{DSAT}$. Below V_{DSAT}, the electric field near the source varies directly with V_{DS}, but above V_{DSAT}, the source electric field increases slowly with increases in V_{DS}. The slow rise in $\langle v(0) \rangle$ with V_{DS} beyond the saturation voltage occurs because of the slow increase in electric field, which slowly decreases r. Since I_{DS} is the product of $\langle v(0) \rangle$ and $Q_i(0)$, which is approximately constant, these observations also explain the I_{DS} versus V_{DS} characteristic displayed in Fig. 9.

B. Carrier Backscattering in a Nano-MOSFET

Given the central role of the backscattering coefficient r in the operation of a MOSFET, we should examine the physics that controls it. The backscattering coefficient is determined by both carrier scattering and by the potential drop within the channel. Fig. 12 schematically illustrates a stream of carriers injected into the channel from the quasi-equilibrium point at the top of the barrier. The fraction that backscatters and returns to the source is defined as r. If backscattering occurs beyond a certain critical distance (denoted as ℓ in Fig. 12), then it is unlikely that the carrier will have sufficient *longitudinal* energy to surmount the barrier and exit into the source. More likely, it will be reflected by the channel potential, perhaps undergo several scattering/electric field reflections, and exit from the drain. These scattering events will increase the carrier density in the channel and through Poisson's equation, the self-consistent electric field throughout the entire channel, but they do not contribute directly to r as we have defined it. To understand why this occurs, one must realize that Fig. 12 is a plot of *longitudinal* energy $(m^* v_x^2/2)$ not *total* energy $(m^* v^2/2)$. For the typical case of a wide MOSFET, there is a continuous distribution of transverse modes. Only a small fraction of the carriers will backscatter directly at the source and possess sufficient longitudinal energy to surmount the barrier. Note that this argument applies to both elastic and inelastic scattering. Finally, note that if this were a quantum wire MOSFET in which the only degree of freedom was the x-axis, then r would be sensitive to backscattering throughout the entire channel.

From the argument presented above, we conclude that the steady-state drain current is limited only by backscattering that occurs within the critical distance ℓ from the beginning of the channel. The existence of such a critical distance was first noted by Price, who observed in performing Monte Carlo simulations of carrier transport down a potential barrier, that if carriers penetrated only a very short distance into the potential drop, even if they did scatter, they were unlikely to return to their injection point at the top of the barrier [28]. Much earlier, Bethe showed that for a forward biased metal-semiconductor junction, currents near the thermionic (i.e., ballistic) limit occur when the first k_BT/q of potential drop at the junction, occurs over a distance much less than the mean-free-path. Since this critical distance (known as the k_BT-layer [15]) is a small fraction of the barrier width, the thermionic emission typically applies. The critical layer for backscattering in a MOSFET is also roughly the distance over which the first k_BT/q of channel potential drops, typically a small fraction of the channel length. This occurs because the longitudinal energy of the backscattered electrons is randomized to have an average energy of k_BT, so that carriers that scatter beyond this point have little chance of returning to the source [33].

By identifying the critical distance ℓ with the k_BT-layer, the expression for the backscattering coefficient for a field free semiconductor slab of length L [10], [25],

$$r = \frac{L}{L+\lambda} \tag{8}$$

where λ is the mean-free-path for backscattering, can be generalized to [5]

$$r = \frac{\ell}{\ell+\lambda}. \tag{9}$$

Since the critical backscattering occurs in a region where the carriers have gained little energy from the channel field, the appropriate mean-free-path to use in (9) is λ_o, the near-equilibrium mean-free-path for backscattering, which can be obtained from the measured mobility of a long-channel MOSFET. A comparison of the simple expression (9), with an evaluation of r by direct Monte Carlo simulation, shows good agreement [5]. Monte Carlo simulations also show that there is no sharp cutoff of scattering at the kT-layer, only that scattering is heavily weighted to the kT-layer. Note also that the key result (9) need not be postulated; it can be derived by scattering theory (see [25, Ch. 9] for an introduction to semiclassical scattering theory).

Calculating the channel backscattering coefficient [even under the simplifying assumptions that lead to (9)] is nontrivial, but a simple argument explains why the importance of backscattering decreases from source to the drain. Consider a charge carrier injected from the source into the channel with momentum $\mathbf{p}_o = (p_{xo}, p_{zo})$, as shown in Fig. 13. (Because of the quantum confinement in the y-direction, the electron has two degrees of freedom). If this injected carrier gains an energy ΔE by acceleration in the longitudinal electric field, then its momentum is $\mathbf{p}_1$, where $p_1^2 = (p_{xo}^2 + 2m\Delta E) + p_{zo}^2$. Assume that the electron then backscatters elastically to momentum $\mathbf{p}_1'$ (see Fig. 13). If the backscattered electron propagates ballistically back to the beginning of the channel, what is the probability that it can cross the barrier, and, therefore,

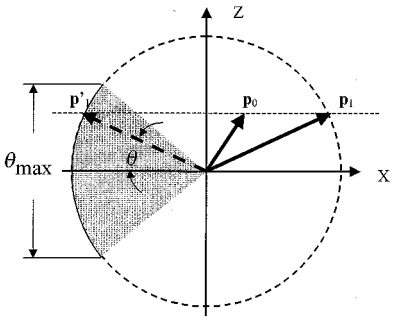


Fig. 13. Illustration of backscattering and how it contributes to r. A 2-D confined carrier is injected into the source with momentum $\mathbf{p_0}$. It propagates down the potential drop toward the drain gaining an energy ΔE with corresponding momentum $\mathbf{p_1}$. It then scatters to momentum $\mathbf{p_1}'$ (since we assume elastic scattering, $\mathbf{p_1} = \mathbf{p_1}'$). Only carriers within the shaded region have sufficient longitudinal momentum to cross the barrier and enter the source.

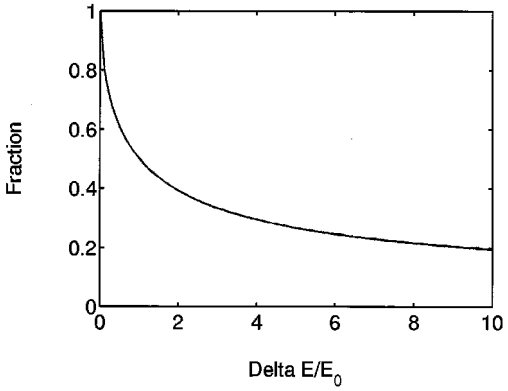


Fig. 14. Fraction of the scattered electrons that contribute to the channel backscattering coefficient, r (i.e., the shaded region in Fig. 13). The curve is evaluated from (12) assuming that a carrier gains an energy, ΔE, before iotropically scattering, then propagates back to the barrier without scattering again.

contribute to r? To do so requires sufficient longitudinal kinetic energy

$$\frac{p_{1x}'^2}{2m^*} = \frac{p_1^2}{2m^*}\cos^2\theta \geq \Delta E. \tag{10}$$

Equation (10) defines a maximum angle $\theta_{\max}$ for backscattered carriers that contribute to r

$$\theta_{\max} = \cos^{-1}\left(\sqrt{\frac{\Delta E}{\Delta E + E_0}}\right) \tag{11}$$

(see Fig. 13). Finally, the fraction of the scattered carriers that contribute to r is the fraction with $|\theta| < \theta_{\max}$ or

$$F = \frac{\theta_{\max}}{\pi} = \frac{\cos^{-1}\left(\sqrt{\frac{\Delta E}{\Delta E+E_0}}\right)}{\pi}. \tag{12}$$

Fig. 14 is a plot of F versus $\Delta E/E_0$; it shows that when the carriers have traveled down the potential drop by an amount equal to the injection energy E_0 (k_BT for a nondegenerate, 2-D carrier gas), then even if they do scatter, only 50% of them have a chance to contribute to r. As carriers travel further down the potential drop, the probability that a scattering event will contribute to the channel backscattering coefficient r steadily decreases.

The simple argument presented above explains why scattering near the source controls the backscattering coefficient r.

In practice, the critical region is even more weighted toward the beginning of the channel than suggested by Fig. 14. There are three reasons why this happens. First, as the backscattered carrier propagates toward the source, it may be scattered again. Second, as the injected carrier penetrates deeper into the channel, its energy increases and so does the probability of scattering by phonon emission, which lowers its energy and makes it less likely to return to the source. Third, surface roughness scattering is strong near the beginning of the channel.

IV. Discussion

Transport in a nanoscale MOSFET is nonlocal; the average carrier velocity does not depend on the local electric field. A mobility can be precisely defined, but since it depends on an essentially unknown distribution function, it is not a useful parameter [29]. Mobility is, however, a well-defined parameter in a long-channel MOSFET. From the near-equilibrium mobility, which is readily measured in a long-channel MOSFET, the near-equilibrium mean-free-path for backscattering, which is the important transport parameter for a nanoscale MOSFET, can be determined. In this sense, one can say that mobility is a meaningful parameter for nanoscale MOSFETs. (There are, of course, complicating factors that have to be dealt with, such as the use of halo implants which can result in different channel dopings for long- and short-channel devices and, therefore, different mobilities).

Shockley used scattering theory to relate the near-equilibrium diffusion coefficient D_o to the mean-free-path for backscattering as [11], [12]

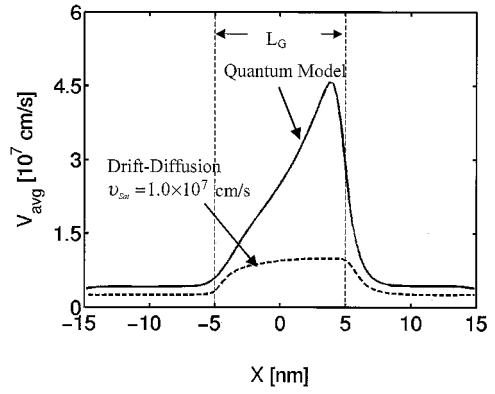
$$D_o = \frac{v_T \lambda_o}{2}. \tag{13}$$

For an alternative derivation of this result, the reader is referred to [25, Ch. 9]. Since near-equilibrium conditions prevail, the Einstein relation may be invoked and the result is a simple relation between the near-equilibrium mobility and the near-equilibrium mean-free-path for backscattering. (The meaning of near-equilibrium is clouded by quasi-ballistic transport; what is required is that each half of the carrier distribution have a near-equilibrium shape). Finally, we note that (13) assumes nondegenerate carrier statistics, but this assumption fails above threshold. In the more general case, the relation between λ_o and μ_o becomes more complex. Note also that defining the width of the critical region from the k_BT/q potential drop also assumes nondegenerate carrier statistics. Our use of nondegenerate statistics establishes the central ideas simply, but it is also clear that degeneracy will be important.

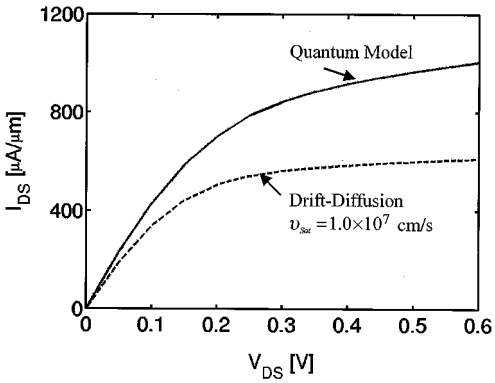
We have been careful to refer to λ_o as the mean-free-path for *backscattering*, but we have not defined it precisely. The relation of the mean-free-path for backscattering that we use and the mean-free-path itself is analogous to the relation between the momentum relaxation time τ_m and the mean time between scattering events τ [25]. This mean-free-path can be precisely define in terms of the transition rate per unit time for scattering from state $\mathbf{k}$ to $\mathbf{k}'S(\mathbf{k},\mathbf{k}')$ as [30]

$$\frac{1}{\lambda_o} \equiv \sum_{k_x'>0,k_z'} \frac{S(\mathbf{k},\mathbf{k}')}{v_x(\mathbf{k})} \tag{14}$$

where we have assumed nondegenerate carrier statistics.



(a)



(b)

Fig. 15. (a) Average velocity versus position at $V_{GS} = V_{DS} = 0.6$ V for the 10-nm DG SOI-MOSFET. Two different transport models are compared. Solid line: Green's function approach that captures velocity overshoot. Dashed line: Drift-diffusion with velocity saturation. (b) I_{DS} versus V_{DS} for $V_{GS} = 0.6$ V for the 10-nm DG-SOI MOSFET. Two different transport models are compared. Solid line: Green's function approach that captures velocity overshoot. Dashed line: Drift-diffusion with velocity saturation.

For the past decade, a lot of the modeling and simulation work has focussed on accurately describing velocity overshoot within the channel, but in the view presented in this paper, velocity overshoot is considered to play an indirect role. It can, however, have significant effects on devices [16]. We should note first that to achieve a velocity at the source that approaches the thermal limit, the velocity within the channel must be even higher. When the source velocity is well below the thermal limit, it is possible for a velocity saturated simulation to get the velocity at the source correct, but it will erroneously clamp the velocity near the drain at an unphysically low value. The inversion layer density in the channel will be too high near the drain, which will lead to errors in the self-consistent channel potential. These carriers will screen the source from charges on the drain, so we should expect a unphysically low output conductance from a velocity-saturated model. These effects are shown in Fig. 15. Fig. 15(a) compares the channel velocity versus position profiles under on-current conditions for a velocity-saturated drift-diffusion transport model and for the Green's function method. We observe that the Green's function method captures the velocity overshoot that occurs near the drain. Fig. 15(b) compares the simulated I_{DS} versus V_{DS} characteristics for the two transport models. Note that the output conductance is considerably higher when velocity overshoot is included. Bude has

observed that the effect can be as much as 40% for nanoscale bulk MOSFETs [16].

V. Summary

A conceptual view of the essential physics of carrier transport in nanoscale MOS transistors was presented and confirmed by 2-D numerical simulation. The key results are as follows.

1) The source velocity saturates and that its limit is set by thermal injection.
2) The carrier density at the top of the source to channel barrier is fixed by MOS electrostatics (in an electrostatically well-designed MOSFET).
3) The scattering in a short region near the beginning of the channel limits the on-current.
4) The role of off-equilibrium velocity overshoot is largely an indirect one based on its influence on the self-consistent potential throughout the channel.

The results show that the physics that determines the steady-state current of a MOSFET can be understood in terms of a simple model. The view of nanoscale MOSFET device physics discussed in this paper (see also [31]) should provide a useful guide for experimental and theoretical work, for developing compact models, and for interpreting detailed simulations.

References

[1] S. E. Laux and M. V. Fischetti, "Monte Carlo simulation of submicrometer Si n-MOSFETs at 77 and 300 K," *IEEE Electron Device Lett.*, vol. 9, pp. 467–469, Dec. 1988.

[2] M. R. Pinto, E. Sangiorgi, and J. Bude, "Silicon transconductance scaling in the overshoot regime," *IEEE Electron Device Lett.*, vol. 14, pp. 375–378, Nov. 1993.

[3] C. Jungemann, S. Keith, M. Bartels, and B. Meinerzhagen, "Efficient full-band Monte Carlo simulation of silicon devices," *IEICE Trans. Electronics*, vol. E82, pp. 870–879, 1999.

[4] D. Javanovic and R. Venugopal, 7th Int. Worklshop Computational Electronics, Glasgow, U.K., May 25–25, 2000.

[5] M. S. Lundstrom, "Elementary scattering theory of the MOSFET," *IEEE Electron Device Lett.*, vol. 18, pp. 361–363, Nov. 1997.

[6] S. Datta, F. Assad, and M. S. Lundstrom, "The Si MOSFET from a transmission viewpoint," *Superlatt. Microstruct.*, vol. 23, pp. 771–780, 1998.

[7] Y. Naveh and K. K. Likharev, "Modeling of 10-nm-scale ballistic MOSFETs," *IEEE Electron Device Lett.*, vol. 21, pp. 242–244, June 2000.

[8] Z. Ren, R. Venugopal, S. Datta, M. S. Lundstrom, D. Jovanovic, and J. G. Fossum, "The ballistic nanotransistor: A simulation study," in *IEDM Tech. Dig.*, Dec. 2000, pp. 715–718.

[9] Y. Taur and T. Ning, *Fundamentals of VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[10] S. Datta, *Electronic Transport in Mesoscopic Systems*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[11] J. P. McKelvey, R. L. Longini, and T. P. Brody, "Alternative approach to the solution of added carrier transport problems in semiconductors," *Phys. Rev.*, vol. 123, pp. 51–57, 1961.

[12] W. Shockley, "Diffusion and drift of minority carrier in semiconductors for comparable capture and scattering mean free paths," *Phys. Rev.*, vol. 125, pp. 1570–1576, 1962.

[13] P. M. Solomon, "A comparison of semiconductor devices for high-speed logic," *Proc. IEEE*, vol. 70, pp. 489–509, 1982.

[14] E. O. Johnson, "The insulated-gate field-effect transistor—A bipolar transistor in disguise," *RCA Rev.*, vol. 34, pp. 80–94, 1973.

[15] F. Berz, "The Bethe condition for thermionic emission near an absorbing boundary," *Solid-State Electron.*, vol. 28, pp. 1007–1013, 1985.

[16] J. Bude, "MOSFET modeling in the ballistic regime," in *Proc. Int. Conf. Simulation of Semiconductor Processes and Devices (SISPAD'00)*, Seattle, WA, Sept. 6–8, 2000, pp. 23–26.

[17] A. Rahman, Z. Ren, J.-H. Rhew, and M. S. Lundstrom, "Toward a compact scattering model for nanoscale MOSFETs," in *Proc. 4th Int. Conf. Modeling and Simulation of Microstructures (MSM'01)*, Hilton Head, SC, Mar. 19–21, 2001.

[18] S. Datta, "Nanoscale device modeling: The Green's function method," *Superlatt. Microstruct.*, vol. 28, pp. 253–278, 2000.

[19] "nanoMOS 2.0," [Online] Available www.ece.purdue.edu/Celab

[20] F. Assad, Z. Ren, S. Datta, M. S. Lundstrom, and P. Bendix, "Performance limits of Si MOSFETs," in *IEDM Tech. Dig.*, Dec. 1999, pp. 547–549.

[21] G. Timp and J. Bude *et al.*, "The ballistic nanotransistor," in *IEDM Tech. Dig.*, Dec. 1999, pp. 55–58.

[22] K. Natori, "Ballistic metal-oxide-semiconductor field effect transistor," *J. Appl. Phys.*, vol. 76, pp. 4879–4890, 1994.

[23] J.-H. Rhew, Z. Ren, and M. Lundstrom, *Numerical Study of a Ballistic MOSFET*, 2001, submitted for publication.

[24] F. Assad, Z. Ren, D. Vasileska, S. Datta, and M. S. Lundstrom, "On the performance limits for Si MOSFETs: A theoretical study," *IEEE Trans. Electron Devices*, vol. 47, pp. 232–240, Jan. 2000.

[25] M. Lundstrom, *Fundamentals of Carrier Transport*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[26] H. Hu, J. B. Jarvis, L. T. Su, and D. A. Antoniadis, "A study of deep-submicron MOSFET scaling based on experiment and simulation," *IEEE Trans. Electron Devices*, vol. 42, pp. 669–677, Mar. 1995.

[27] J. Bude, private communication, Dec. 1999.

[28] P. J. Price, "Monte Carlo calculation of electron transport in solids," *Semicond. Semimet.*, vol. 14, pp. 249–334, 1979.

[29] S. Bandyopadhyay, C. M. Maziar, M. E. Klausmeier-Brown, S. Datta, and M. S. Lundstrom, "Rigorous technique to couple Monte Carlo and drift-diffusion models for computationally efficient device simulation," *IEEE Trans. Electron Devices*, vol. ED-34, pp. 392–399, Feb. 1987.

[30] M. A. Alam, M. A. Stettler, and M. S. Lundstrom, "Formulation of the Boltzmann equation in terms of scattering matrices," *Solid-State Electron.*, vol. 36, pp. 263–271, 1993.

[31] K. Natori, "Scaling limit of the MOS transistor—a ballistic MOSFET," *IEICE Trans. Electron.*, vol. E84-C, pp. 1029–1036, 2001.

[32] M. S. Lundstrom, Z. Ren, and S. Datta, "Essential physics of carrier transport in nanoscale MOSFETs," in *Proc. Int. Conf. Simulation of Semiconductor Processes and Devices (SISPAD'00)*, Seattle, WA, Sept. 6–8, 2000, pp. 1–5.

[33] J. H. Rhew, private communication, Sept. 2001.

Mark Lundstrom (S'72–M'74–SM'80–F'94) received the B.S. and M.S. degrees from the University of Minnesota, Minneapolis, and the Ph.D. degree from Purdue University, West Lafayette, IN.

Currently, he is Professor of Electrical and Computer Engineering at Purdue University, where he has also served as Assistant Dean of Engineering and as Director of the Optoelctronics Research Center. Before joining Purdue, he was with Hewlett-Packard Corporation, where he worked on the development of an integrated circuit manufacturing process. His teaching and research now centers on the physics, technology, and simulation of nanoscale electronic devices.

Dr. Lundstrom is a Fellow of the IEEE and APS. His work in education has been recognized by the Frederick Emmons Terman Award from the American Society of Engineering Education (ASEE). For his work on nanoscale electronic devices, he was the corecipient (with S. Datta) of IEEE's 2002 Cledo Brunetti Award.

Zhibin Ren was born in Zhengzhou, China. He received the B.S. degree in physics from Zhejiang University, China, in 1991, and the M.S. degree in physics from the University of Massachusetts at Dartmouth in 1997. In 1997, he joined the Device Simulation Group in the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, where he is currently pursuing the Ph.D. degree. His research interests are primarily centered on device physics, modeling, and simulation of nanoscale transistors.