Compact Models and the Physics of Nanoscale FETs

Mark S. Lundstrom, *Fellow, IEEE*, and Dimitri A. Antoniadis, *Fellow, IEEE*
(Invited Paper)

***Abstract*— The device physics of nanoscale MOSFETs is related to traditional compact models. Beginning with the virtual source model, a model for nanoscale MOSFETs expressed in traditional form, we show how the Landauer approach gives a clear physical interpretation to the parameters in the model. The analysis shows that transport in the channel is limited by diffusion near the virtual source both below and above threshold, that the current saturation is determined by velocity saturation near the source, not by the maximum velocity in the channel, and that the channel resistance approaches a finite value as the channel length approaches zero. These results help explain why traditional models continue to work well at the nanoscale, even though carrier transport is distinctly different from that at the microscale, and they identify the essential physics that physics-based compact models for nanoscale MOSFETs should comprehend.**

***Index Terms*— Ballistic transport, MOSFETs, nanoelectronics, semiconductor device modeling.**

I. Introduction

PHYSICS-BASED compact models for electronic devices play two important and distinct roles. First, the kernel of the model serves as a compact mathematical description of our understanding of the device. This conceptual model helps us to interpret experiments and detailed simulations and guides our thinking in device research and development. Second, the complete model with extensions to treat parasitic elements and with consideration of the practicalities that ensure that it runs robustly in a circuit simulator [1] enables circuit design. Today's sophisticated compact models for field-effect transistors (FETs) (e.g. [2]–[7]) have evolved from models first developed 30–50 years ago [8]–[11]. For much of the physics, this is appropriate; the Poisson equation is still valid, but carrier transport at the nanoscale is different.

In nanoscale FETs, velocity overshoot and ballistic and quasi-ballistic transport become important [12]–[15]. Typically, these effects are treated as perturbations to the traditional approach, but they are not perturbations; transport in a nanoscale FET is distinctly different from transport in microscale FETs [12]–[16]. In fact, the whole premise of the gradual channel approximation on which all FET models are built should be reexamined under near-ballistic conditions [17]. (It should hold, however, very near the beginning of the channel, which is the key to the model to be discussed). Surprisingly, traditional FET compact models continue to fit well the current–voltage characteristics of nanoscale devices if the mobility and saturation velocity are treated as fitting parameters. On the other hand, the magnitude of channel charge versus voltage is demonstrably overestimated [17]. The use of empirical models may be acceptable for circuit simulation, particularly in the presence of large parasitic and load capacitances, but it causes confusion when using them as conceptual guides for device research and development.

Our goal in this paper is to relate the physics of nanoscale MOSFETs to the traditional theory used for compact models. This paper reviews the understanding of nanoscale MOSFETs that has been developed through computational and experimental studies over the past 15 years. We also relate this physical understanding to the virtual source (VS) compact model [18]. The approach is tutorial because we aim to convey this understanding to those in the compact modeling community who have not closely followed this work.

This paper continues in Section II by discussing the MOSFET in terms of energy band diagrams, which makes the essential physics clear and highlights the similarity of FETs and bipolar transistors. In Section III, we present a simple version of the VS model [18] and relate it to the barrier-controlled approach of Section II. Subsequent discussions clarify the physical basis of two key parameters in the VS model, the apparent mobility and the injection velocity. In Section IV, we develop a MOSFET model using the Landauer–Boltzmann approach to carrier transport. (This approach treats ballistic and quasi-ballistic transport in nanoscale devices and reduces to the drift-diffusion equation for long devices [19], [20]). The Landauer model for the MOSFET looks much different from the traditional model, but we show in Section V that it can be reexpressed in the traditional form and that doing so gives a clear physical meaning to parameters in the VS model. This paper concludes in Section VI by summarizing the challenges of developing compact models that fully capture the essential physics of modern FETs.

II. Transistor as a Barrier Controlled Device

Fig. 1(a) is a sketch of the equilibrium conduction band edge versus position along the surface of an N-channel MOSFET

Manuscript received June 27, 2013; revised August 23, 2013; accepted September 18, 2013. Date of publication October 9, 2013; date of current version January 20, 2014. This work was supported by the National Science Foundation through the NCN-NEEDS program under Contract 1227020-EEC. The review of this paper was arranged by Editor X. Zhou.

M. S. Lundstrom is with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: lundstro@purdue.edu).

D. A. Antoniadis is with the Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: daa@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2013.2283253

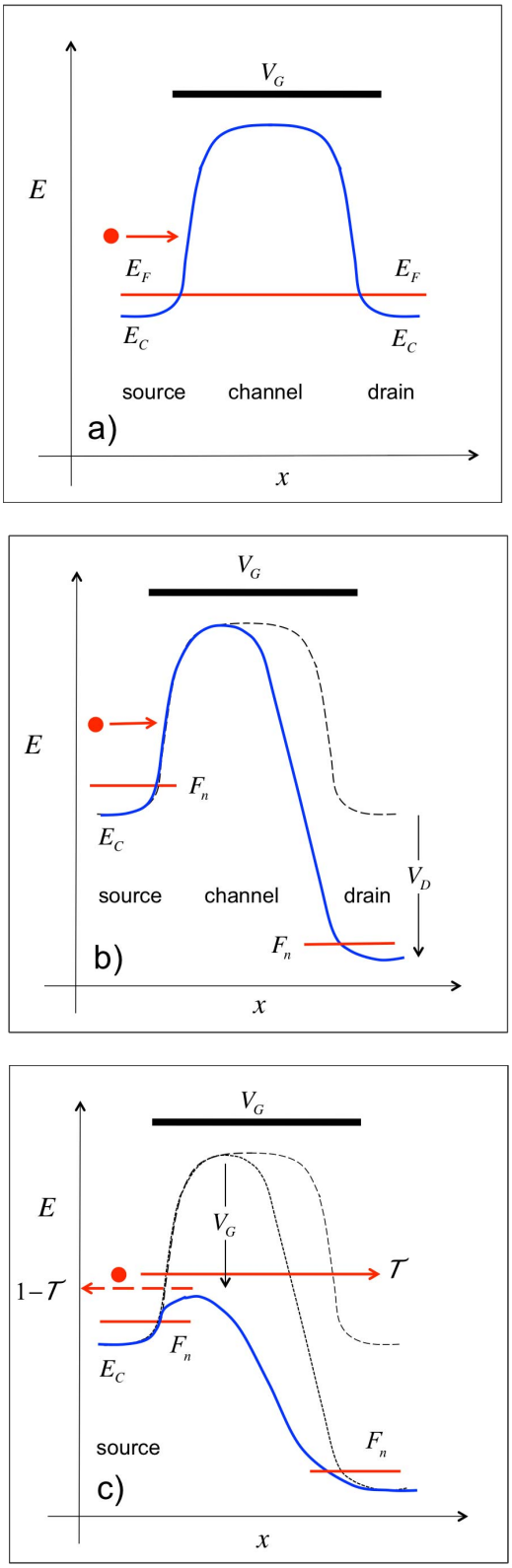


Fig. 1. Operation of a MOSFET in terms of energy band diagrams. (a) Equilibrium conduction band-edge versus position for a low gate voltage. (b) Conduction band-edge E_C versus position for a large applied drain bias. (c) Conduction band-edge versus position for a large gate and drain bias. (F_n is the electron quasi-Fermi level).

from the source, across the channel, to the drain. Because the channel is p-type (or undoped), there is a large barrier that prevents electrons in the source from flowing to the drain. Note the similarity of this energy band diagram to that of an NPN bipolar transistor.

As shown in Fig. 1(b), the application of a large drain voltage lowers the quasi-Fermi level in the drain and, along with it, the conduction band edge. If, however, the MOSFET is electrostatically well designed, then the height of the energy barrier is primarily controlled by the gate. The device is off, and there is only leakage current due to electrons in the source that are thermionically emitted over the source to channel barrier.

As shown in Fig. 1(c), when a large gate voltage is also applied, the potential in the channel increases, which lowers the energy barrier to the source. Thermionic emission over the source to the channel barrier increases exponentially, and the device turns on. For low gate voltages (subthreshold), the surface potential varies with the gate voltage as $\psi_S = V_{\rm GS}/m$, where $m \approx 1$ [6], [7]. Because the probability for thermionic emission increases exponentially as the barrier is lowered, the subthreshold current varies exponentially with gate voltage. Above threshold, electrons in the channel screen the potential from the gate, so the surface potential varies logarithmically with the gate voltage. The drain current still depends exponentially on the barrier height, so $I_{\rm DS}$ varies linearly with gate voltage for $V_{\rm GS} > V_T$. Both the drain current of a MOSFET and the collector current of a bipolar transistor are controlled by diffusion-like transport over a barrier near the source or emitter [21]. For the MOSFET, the height of the barrier is controlled by the gate to source voltage, and for the bipolar transistor by the base to emitter voltage.

The on-state operation of a well-designed MOSFET can be understood from Fig. 1(c). Because the device is electrostatically well designed, there is a short region near the source where the potential is mostly controlled by the gate voltage, and the lateral electric field is small. Increases in the drain voltage have a small effect on this region if the transistor's drain-induced barrier lowering (DIBL) is low, so the current is relatively insensitive to increases in the drain voltage. Drain current saturation occurs because of 2-D electrostatics, but what controls the magnitude of the current?

As shown in Fig. 1(c), a flux of electrons is thermionically emitted over the barrier, and into the channel. A fraction T exits from the drain, and a fraction $1 - T$ backscatters and returns to the source. At the top of the barrier, the lateral electric field is zero. The goal of transistor design is to make the charge at the top of the barrier (the VS) equal to value given by ideal 1-D MOS electrostatics with only a small correction due to DIBL (i.e., $Q_n(0) = C_{\rm inv}\ (V_{\rm GS} - V_T)$, where $Q_n(0)$ is the inversion charge at the top of the barrier, $C_{\rm inv}$ is the gate capacitance measured under strong inversion conditions, $V_T = V_{T0} - \delta V_{\rm DS}$ is the threshold voltage with V_{T0} being the threshold voltage at an infinitesimally small drain to source voltage $V_{\rm DS}$, and δ is the DIBL).

The MOSFETs drain current per unit width is proportional to the product of the charge density and velocity. Since the current is constant (no recombination), it can be evaluated anywhere, but the VS is the best choice because the gradual channel approximation applies there, so we know the inversion

layer charge there. Accordingly

$$I_{\mathrm{DS}} = W Q_n(0) \langle v(0) \rangle . \tag{1}$$

The velocity at the beginning of the channel (at the VS) is $\langle v(0) \rangle$, which is known as the injection velocity. Electrons injected from the source diffuse across the low-field region, and when they enter the high-field part of the channel, they are swept across and out through the drain. Diffusion cannot occur faster than the thermal velocity; when the diffusion velocity in the low-field region reaches the thermal (ballistic injection) velocity, the transistor is operating at its ballistic limit. The low-field part of the channel is like the base of a bipolar transistor, and the high-field part is like the collector [21]. In contrast to a long-channel device, where transport is by diffusion below threshold and by drift above threshold, it is by diffusion near the VS in both cases for a nanoscale MOSFET operating below the ballistic limit.

Our discussion has described the MOSFET saturation region in terms of energy band diagrams, but similar arguments apply to the linear region [16]. Understanding the MOSFET as a barrier-controlled device omits many details, some of which can become important in certain cases, but it seems to describe the essential physics of MOSFETs with nanoscale channel lengths and is a useful guide to the development of compact models. Several detailed computational studies support the general picture described here [22]–[30], although others identify additional complexities [31]–[34].

The art of compact modeling is the balance between rigorous and empirical modeling. The model described here seems to provide a good conceptual foundation for understanding small FETs. It shows that the shape of the MOSFET I–V characteristic is determined by manipulating the energy barrier with the gate and drain voltages, i.e., by 1-D and 2-D electrostatics, which becomes more challenging at the nanoscale, but does not fundamentally change. The magnitude of the current, however, is determined by transport, which is much different at the nanoscale than at the microscale.

III. Virtual Source Model

The VS model is a simple semiempirical compact model for MOSFETs that describes a wide range of nanoscale FETs with good accuracy [18]. Although the VS model was originally developed directly from the point of view of quasi-ballistic transport, which we will adopt later, it is useful in this section to present a quick development of the basic model following the traditional MOSFET theory. We will then show that by simply reinterpreting the effective mobility and the saturation velocity, nanoscale FETs can be accurately modeled. The physical justification for this reinterpretation will be discussed in Section IV.

Above threshold and under low drain-to-source bias, the electric field in the channel is approximately V_{DS}/L, and the drift velocity is $\mu_{\mathrm{eff}} V_{\mathrm{DS}}/L$, where μ_{eff} is the so-called effective mobility of traditional MOSFET theory, i.e., the depth-averaged mobility in the channel [6], [7].

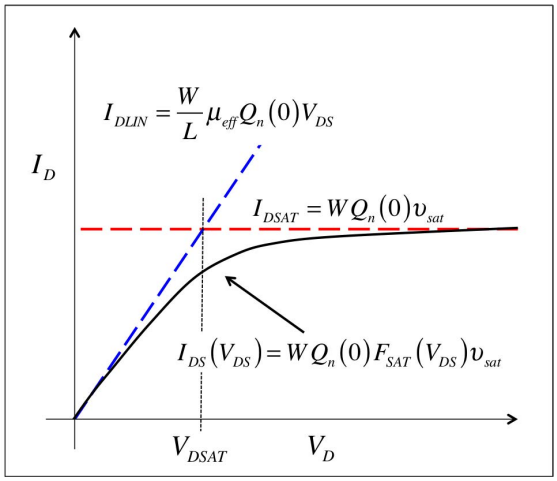


Fig. 2. Illustration of how the virtual source VS model empirically combines expressions for the linear and saturation region currents to obtain a full range of I–V characteristic.

Accordingly, we can write the drain current (1) as

$$I_{\mathrm{DLIN}} = W Q_n(0) \langle v(0) \rangle = \frac{W}{L} \mu_{\mathrm{eff}} Q_n(0) V_{\mathrm{DS}} \tag{2a}$$

which is the standard expression for the linear region current.

In a nanoscale MOSFET under high drain bias, the longitudinal electric field is very high. (For a 30-nm channel length with 1 V on the drain, the average electric field is over 300 kV/cm). In bulk silicon, the electron velocity saturates at about $v_{\mathrm{SAT}} \approx 10^7$ cm/s for electric fields above 10 kV/cm, so one might assume that the electron velocity is saturated along the entire channel. Accordingly, (1) gives the saturation current as

$$I_{\mathrm{DSAT}} = W Q_n(0) \langle v(0) \rangle = W Q_n(0) v_{\mathrm{SAT}}. \tag{2b}$$

As shown in Fig. 2, (2a) and (2b) give the current at low and high drain voltages; the two currents intersect at a voltage given by

$$V_{\mathrm{DSAT}} = \frac{v_{\mathrm{SAT}} L}{\mu_{\mathrm{eff}}}. \tag{3}$$

(The full VS model includes a refinement to properly treat saturation in weak inversion [18].) The current for an arbitrary drain voltage is the smaller of the two expressions. The VS model takes an empirical approach and defines the full range I–V characteristics using a saturation function that smoothly increases the velocity from $\mu_{\mathrm{eff}} V_{\mathrm{DS}}/L$ to v_{SAT} as the drain voltage increases. The VS model I–V characteristic is given by

$$I_{\mathrm{DS}} = W Q_n \left(V_{\mathrm{GS}}, V_{\mathrm{DS}} \right) F_{\mathrm{SAT}} \left(V_{\mathrm{DS}} \right) v_{\mathrm{SAT}} \tag{4a}$$

where

$$F_{\mathrm{SAT}} \left(V_{\mathrm{DS}} \right) = \frac{\left(V_{\mathrm{DS}} / V_{\mathrm{DSAT}} \right)}{\left(1 + \left(V_{\mathrm{DS}} / V_{\mathrm{DSAT}} \right)^{\beta} \right)^{\frac{1}{\beta}}} \tag{4b}$$

and $Q_n \left(V_{\mathrm{GS}}, V_{\mathrm{DS}} \right)$ is given by the MOS electrostatics. A numerical solution based on the surface potential could be

used for Q_n, but the VS model uses an empirical function that smoothly goes from subthreshold to above threshold [18]. This expression comprehends 1-D and 2-D electrostatics through the measured subthreshold slope and DIBL.

As we have described it, the VS model is a traditional MOSFET model with only a few physical parameters, namely gate capacitance in inversion, subthreshold swing, DIBL, and the threshold voltage or off-current. With these parameters as inputs, fitting the model to experimental data determines the mobility, velocity, and series resistance. The VS model generally produces excellent fits to measured data, but two things are noticed. First, the extracted mobility tends to decrease with channel length, especially in III–V FETs, which are not clouded by halos and other effects that could increase scattering at short channel lengths [35]. This is not unexpected because (2a) does not behave well as $L \to 0$ where it should approach a finite ballistic limit [36], [37]. The mobility must decrease as $L \to 0$ so that (2a) approaches a finite limit, but the physical interpretation of this mobility is not clear. Second, the meaning of the extracted saturation velocity obtained from curve fitting is not clear. How does it relate to the high-field bulk saturation velocity, to the peak of the velocity versus field characteristic in a III–V material, and to the velocity overshoot, which is known to occur in short high-field regions [12]–[14], [38]?

In the next section, we will derive by a completely different method a simple expression for the I–V characteristic of a nanoscale MOSFET. When we compare this new expression to the VS model, a clear physical interpretation will emerge.

IV. Landauer Approach to Nano-Scale FETs

To treat transport at the nanoscale, we express the drain current as [19], [20], and [39]

$$I_{\rm DS} = \frac{2q}{h}\int \mathcal{T}(E)\, M(E)\,(f_S - f_D)\, dE \tag{5}$$

where $\mathcal{T}(E)$ is the transmission at energy E, $M(E)$ is the number of current-carrying channels at energy E, and $f_S(E)$ and $f_D(E)$ are the equilibrium Fermi functions in the source and drain. Scattering near the VS is nearly elastic, but in the high-field portion of the channel it is strongly inelastic. Equation (5) applies when the scattering is elastic, so we use it near the VS. Although it is sometimes thought that the Landauer approach applies only to the ballistic transport, but, as discussed in [19] and [20], it applies from the ballistic to diffusive regimes.

For a derivation of (5) and a discussion of the underlying assumptions, see [19] and [20], but the result is readily understood. For uniform temperature, $f_S(E)$ and $f_D(E)$ differ when a voltage on the drain lowers the drain Fermi level with respect to the source; so (5) states that the current is zero when the voltage across the device is zero. Equation (5) also states that the current is proportional to the number of current-carrying channels and to the transmission, which is the probability that an electron injected from the source exits from the drain. In the ballistic limit, $\mathcal{T}(E) = 1$. Equation (5) does not resolve quantities spatially (for that we need the Boltzmann equation), so it must be applied to a single location within the device. The obvious location is at the top of the barrier (or virtual source) where the gradual channel approximation holds, and the electron charge $Q_n(0)$ is known.

Consider, first, the linear region (small $V_{\rm DS}$, i.e., $V_{\rm DS} \ll k_BT/q$) where $f_S \approx f_D$. Using a Taylor series expansion for $(f_S - f_D)$, (5) becomes

$$I_{\rm DLIN} = \frac{2q^2}{h}\left\{\int \mathcal{T}(E)\, M(E)\left(-\frac{f_0}{\partial E}\right) dE\right\} V_{\rm DS}. \tag{6}$$

To evaluate this expression, we first assume ballistic transport, $\mathcal{T}(E) = 1$, and nondegenerate carrier statistics so that $f_0 = \exp[(E_F - E)/k_BT]$. (The nondegeneracy assumption is not valid above threshold, but MOS theory typically uses this assumption, so it will make comparisons easier). To proceed, we must specify the number of channels, $M(E)$, which is directly related to the band structure [19], [20]. For a simple parabolic band with an effective mass m^* and valley degeneracy g_V, it is straightforward to show [20]

$$M(E) = g_V W \frac{\sqrt{2m^*(E - E_C)}}{\pi\hbar}. \tag{7}$$

(The parameter W is the width of the MOSFET and g_V is the valley degeneracy.) With these assumptions, we can evaluate (6) to find

$$I_{\rm DLIN} = W\, Q_n\left(V_{\rm GS}, V_{\rm DS}\right)\frac{\upsilon_T}{2\left(k_BT/q\right)} V_{\rm DS} \tag{8}$$

where

$$\upsilon_T = \sqrt{2k_BT/\pi m^*} \tag{9a}$$

is the unidirectional thermal velocity and

$$\begin{aligned} Q_n = qn_S &= \left(g_V\frac{m^*}{\pi\hbar^2}k_BT\right)e^{(E_F - E_C)/k_BT} \\ &= N_{2D}e^{(E_F - E_C)/k_BT}. \end{aligned} \tag{9b}$$

Equation (8) should be compared with the corresponding traditional expression (2a).

Next, we evaluate the current for large $V_{\rm DS}$, where $(f_S - f_D) \approx f_S$. In this case, the drain current expression (5) becomes

$$I_{\rm DSAT} = \frac{2q}{h}\int \mathcal{T}(E)\, M(E) f_S(E) dE. \tag{10}$$

Assuming ballistic transport and proceeding as before, we find

$$I_{\rm DSAT} = W\, Q_n\left(V_{\rm GS}, V_{\rm DS}\right)\upsilon_T \tag{11}$$

which should be compared with to the traditional expression, (2b).

Whereas III–V FETs operate close to the ballistic limit [35], [40], and [41], Si MOSFETs typically operate at about one-half the ballistic limit [42]–[45], so we cannot assume that $\mathcal{T}(E) = 1$. A simple expression for $\mathcal{T}(E)$ in a uniform field-free region can be derived from the Boltzmann equation as [19] and [20]

$$\mathcal{T}_{LIN}(E) = \frac{\lambda(E)}{\lambda(E) + L} \tag{12}$$

where $\lambda(E)$ is the mean free path for backscattering. Equation (12) shows that $\mathcal{T} \to 1$ when $\lambda \gg L$ (ballistic limit) and $\mathcal{T} \to \lambda/L$ when $\lambda \ll L$ (diffusive limit).

If we assume (to keep the math simple) that λ is energy independent; then to obtain the linear current in the presence of scattering we simply multiply (8) by λ to find

$$I_{\mathrm{DLIN}} = T_{LIN} W Q_n \left(V_{\mathrm{GS}}, V_{\mathrm{DS}}\right) \frac{v_T}{2\left(k_B T/q\right)} V_{\mathrm{DS}}. \tag{13}$$

To treat scattering under high drain bias, MOS electrostatics must be considered. Under ballistic conditions, there is a positively directed flux injected from the source and a negatively directed flux injected from the drain. For a drain voltage of a few $k_B T/q$, thermionic emission from the drain to the top of the barrier is suppressed (i.e., $f_S \gg f_D$), and there is only a positively directed net flux at the top of the barrier. As shown in Fig. 1(c) for high drain bias, in the presence of scattering, there are both positive and negative moving fluxes of carriers at the top of the barrier even under high drain bias. MOS electrostatics demands that the charge be the same whether or not there is a backscattered flux. By enforcing MOS electrostatics according to the argument of [46], we find that the high V_{DS} drain current, i.e., (11), must be multiplied by $\mathcal{T}/(2-\mathcal{T})$ to find

$$I_{\mathrm{DSAT}} = \left(\frac{\mathcal{T}_{\mathrm{SAT}}}{2-\mathcal{T}_{\mathrm{SAT}}}\right) W Q_n \left(V_{\mathrm{GS}}, V_{\mathrm{DS}}\right) v_T. \tag{14}$$

To summarize, we have derived expressions for the linear and saturation region currents by applying the Landauer formula at the top of the barrier. The results, i.e., (13) and (14), are the counterparts of the corresponding results from the traditional approach (2a) and (2b). It should be noted that we have labeled the transmission differently in the linear and saturation regions. In the linear region, the transmission is given by (12). An electron scattering anywhere in the channel has a chance to return to the source. Under high bias, however, only the low-field region near the beginning of the channel matters. If electrons transmit across this region and enter the high-field part of the channel, then, even if they scatter, they are unlikely to return to the source [16], [46]. Accordingly, the relevant length is not the whole channel but only the low-field part, and the transmission becomes

$$\mathcal{T}_{\mathrm{SAT}}(E) = \frac{\lambda(E)}{\lambda(E) + \ell} \tag{15}$$

where ℓ is the length of the low-field portion of the channel. Since $\ell \ll L$, the transmission is significantly larger for large V_{DS} than for small V_{DS}. Even though the large drain voltage increases the electron energy so that electrons scatter more, the additional scattering takes place in a part of the channel where it matters little. The device is less ballistic under high drain bias, but the fraction of carriers that transmit from the source to the drain is higher than under low V_{DS}.

V. Discussion

Two different analytical approaches to the MOSFET have been discussed—a traditional approach, which accurately describes nanoscale MOSFETs if we regard the mobility and saturation velocity as fitting parameters; and the Landauer approach, which has clear physical meaning at the nanoscale. The traditional and Landauer expressions look very different, but we will show that the Landauer expressions can be written in a form that is very similar to the traditional form; doing so provides a clear physical interpretation for the mobility and saturation velocity in the VS model.

Consider the linear region current first. The Landauer approach is expressed in terms of the transmission, which involves the near-equilibrium mean free path for backscattering. There is a simple relation between the mean free path and the diffusion coefficient in the bulk [19]: $D_n = v_T \lambda_0/2$, where we assume an energy-independent mean free path λ_0. The near-equilibrium diffusion coefficient in the bulk is related to the mobility by the Einstein relation, so we can relate the effective mobility of electrons in the inversion layer to the mean free path as

$$\mu_{\mathrm{eff}} = \frac{v_T \lambda_0}{2k_B T/q}. \tag{16}$$

(The use of a concept like mobility in a channel that may be shorter than the mean free path can be questioned. In this context, mobility should be understood as simply another way to write the near-equilibrium mean free path.)

Using (12) and (16), we can rewrite (13) as

$$I_{\mathrm{DLIN}} = \frac{W}{L} \mu_{\mathrm{app}} Q_n \left(V_{\mathrm{GS}}, V_{\mathrm{DS}}\right) V_{\mathrm{DS}} \tag{17}$$

where we have defined an apparent mobility by

$$\frac{1}{\mu_{\mathrm{app}}} \equiv \frac{1}{\mu_{\mathrm{eff}}} + \frac{1}{\mu_B}. \tag{18}$$

The term μ_B in (18) is the so-called ballistic mobility [47], [20]

$$\mu_B \equiv \frac{v_T L}{2k_B T/q} \tag{19}$$

and μ_{eff} is the effective mobility of inversion layer electrons due to scattering in the channel as given by (16). In the expression for the ballistic mobility, i.e., (19), we simply replace the actual mean free path in (16) with the channel length. The physical interpretation is that, in a ballistic FET, carriers are thermalized in the source and drain, i.e., they scatter frequently in both these regions, whereas they travel ballistically across the channel, so the distance between scattering events is the channel length itself.

According to (18), the apparent mobility of a MOSFET is the smaller of the real mobility and the ballistic mobility. Since the ballistic mobility is proportional to the channel length, it is most important for very short channels and for very high effective mobilities. Experimentally, one finds that the apparent mobility deduced for an FET decreases as the channel length decreases, even where there is no increase in scattering. The effect is large enough to be measurable for Si MOSFETs [48] and quite distinct in III–V HEMTs [35], [40]. For very short channels, the ballistic mobility dominates, and the linear region current becomes independent of channel length with a value given by Natori's ballistic limit [36], [37].

The Landauer analysis shows that the mobility in the VS model is the apparent mobility as given by (18). What about the saturation current? Using (15) in (14), we find

$$I_{\mathrm{DSAT}} = W Q_n \left(V_{\mathrm{GS}}, V_{\mathrm{DS}}\right) v_{\mathrm{inj}} \tag{20}$$

where the injection velocity is given by

$$v_{inj} = \left[\frac{1}{v_T} + \frac{1}{\left(D_n/\ell\right)}\right]^{-1}. \tag{21}$$

The injection velocity is seen to be the smaller of the velocity at which electrons diffuse across the low-field region at the beginning of the channel, D_n/ℓ, and the velocity v_T at which they are thermionically emitted across the barrier into the channel. When the low-field region is very short, the injection velocity reaches its ballistic limit v_T. We see that the saturation velocity of the traditional approach is actually the injection velocity at the VS.

By replacing the mobility in the traditional model with a well-defined apparent mobility and the high-field bulk saturation velocity with the injection velocity at the VS, the VS model accurately describes nanoscale MOSFETs. The VS model displays the signature of velocity saturation (drain current increasing linearly with gate overdrive), but the physics is much different. In a long high-field region, the velocity saturates because of the steady-state balance between the accelerating force of the electric field and collisional dissipation. In a nanoscale MOSFET, the velocity saturates at the top of the barrier for completely different reasons.

In the ballistic case, drain current saturation occurs when $f_S \gg f_D$, which occurs when V_{DS} is greater than a few $k_B T/q$. For drain voltages of this magnitude, the thermionic emission of electrons from the drain to the top of the barrier is suppressed. For a MOSFET operating below the ballistic limit, drain current saturation occurs because of 2-D electrostatics. In a well-designed MOSFET, the length of the low-field region, ℓ, varies only slowly with drain bias, so the injection velocity as given by (21) is relatively independent of V_{DS}. Note that the VS model is consistent with the notion that the carriers' velocity increases as they move toward the drain. If the carriers are ballistic and their energy profile is known, their velocity can be calculated exactly and therefore the total channel charge can be calculated self-consistently with the current. This idea has been used to allow for proper charge calculation in the VS model even under near-ballistic conditions [17].

We have shown that a MOSFET model expressed in traditional form can accurately describe the I–V characteristics of nanoscale FETs if two key model parameters, the mobility and saturation velocity, are reinterpreted. The VS model is, however, semiempirical and not predictive. For example, the injection velocity must be determined by fitting measurements or detailed simulations, because it requires knowledge of ℓ, which is difficult to compute. The VS model makes the transition from linear to saturation with an empirical saturation function; a proper treatment would require us to model the variation of the length of the critical region from L to ℓ as V_{DS} increases. Finally, note that the VS model is a short-channel model and does not treat long-channel MOSFETs for which $I_{\mathrm{DS}} \propto (V_{\mathrm{GS}} - V_T)^2$.

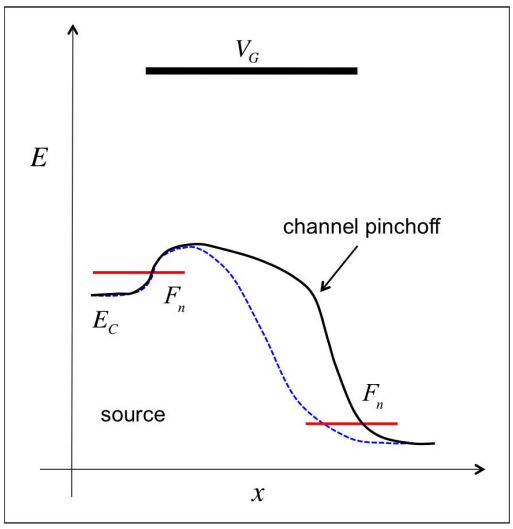


Fig. 3. Sketch of the conduction band versus. position along the channel for the on-state of a long-channel MOSFET. The dashed line is the corresponding result for a short-channel device [i.e., for the nanoscale MOSFET shown in Fig. 1(c)].

It is interesting to compare the on-state energy band diagrams of long- and short-channel MOSFETs. The short-channel case was shown in Fig. 1(c). For a long channel, there is considerable inversion layer density over a substantial portion of the channel, as shown in Fig. 3. In the region of significant electron density, carrier transport is by drift in a nonuniform electric field. For a large enough drain bias, the electron density near the drain drops to a low value, which increases the local resistance of the channel and results in most of the additional drain voltage being dropped across this pinched-off region [9]. Standard velocity saturation models for MOSFETs clamp the velocity to v_{SAT} in this high-field pinched-off region.

Traditional MOSFET theory treats the on-state from short channels where $I_{\mathrm{DS}} \propto (V_{\mathrm{GS}} - V_T)^1$ to long channels where $I_{\mathrm{DS}} \propto (V_{\mathrm{GS}} - V_T)^2$ by invoking velocity saturation in high electric fields. Our best understanding of electron transport, however, shows that there is strong velocity overshoot and no velocity saturation in short high-field regions [12]–[16]. Even in a long-channel MOSFET, the length of the pinched-off region is short, so it is likely that strong velocity overshoot occurs always in the pinch-off region. A treatment of the linear to square law characteristic based on a Buttiker probe approach has been reported [49], [50], but how to do this in a VS context is still not clear.

The model for the nanoscale MOSFET outlined in this paper contains several simplifications, some necessary and some not. For example, we assumed nondegenerate carrier statistics and a simple band structure; these assumptions were made to simplify the analysis and facilitate comparison with traditional MOSFET models. The location of the VS was assumed to be at the top of the barrier, but careful analysis shows that it can be located a short distance after the top of the barrier [51]. In the VS model, the inversion capacitance is measured separately with an MOS capacitor. In a MOSFET,

it should depend on the drain bias as well as gate bias; this effect is not included in the present version of the VS model because its effect is not clearly observed in experimental data. We have not resolved the charge along the channel, so the effects of scattering near the drain, which increases the charge in the channel and electrostatically couples to the potential at the VS, has not been treated. More detailed device models are needed to treat this effect, but we find the effect to be too small to observe in electrostatically well-designed MOSFETs.

Another assumption made in this paper is that of a perfect source, i.e., one that can supply any charge $Q_n(0)$ to the top of the barrier that MOS electrostatics demands. In practice, devices may suffer from source exhaustion, where the source doping is not heavy enough to supply carriers to the top of the barrier [52], or source starvation, in which scattering in the source is not strong enough to keep the momentum states that inject carriers into the channel filled [34], [53]. Both these effects would manifest themselves as a nonlinear (i.e., current dependent) source resistance. Note also that FETs made with novel channel materials often suffer from poor source/drain contacts with high resistance and frequently do not have heavily doped source/drain extensions. Such devices are source-limited and may not follow the behavior outlined here, though the VS model can be adapted to comprehend these effects [54]; this topic falls outside the scope of this paper.

We set out in this paper to explain why traditional MOSFET theory, originally developed for microscale MOSFETs, continues to describe the performance of the smallest transistors being currently manufactured. It can be said that this question is a tautology [54]. The I–V characteristics of a MOSFET are what define the device. Traditional models were developed to explain these characteristics. Several decades of device scaling have preserved the shape of these characteristics; if they had changed significantly, we would not call the device a transistor. Since the traditional model describes I–V characteristics of this shape, it must be possible to fit the model to today's devices. Although this is true, we have shown that the connection between the traditional microscale model and the Landauer nanoscale model is much deeper. The main reason is that the common electrostatics they share is what determines the shape of the I–V characteristics. When the results of the Landauer analysis are expressed in the traditional form, they provide a clear physical interpretation of the parameters in the traditional model when applied to nanoscale MOSFETs.

VI. Conclusion

Our goal has been to provide a clear and succinct description of how we understand carrier transport in nanoscale FETs and to relate this understanding to the widely used traditional model of FETs. The model we have described omits many details, some of which can be important in practice, but we believe that it provides a sound starting point for understanding small MOSFETs. More discussion of the topics raised in this paper is available online [56] as is the VS model itself [57].

Finally, one might ask what a fully physical model for an FET would look like. It would accurately describe MOS electrostatics, as current models do. It would describe drain current saturation from pinch-off in long-channel devices to saturation at the source in nanoscale FETs. It would gracefully approach the ballistic limit as the channel length approaches zero or the mobility approaches infinity. The development of such a model is a worthy intellectual challenge, but is such a model really needed?

From a compact modeling perspective, a strong connection to physics is not essential for circuit simulation (although there are several advantages to physics-based models [7, p. 601]). It is true, however, that the effects discussed here are becoming large enough to measure in Si MOSFETs, are even more pronounced in III–V FETs, and could be even more important in some of the novel channel materials now being investigated. It does not seem likely, however, that traditional MOSFET models will be supplanted by more physical models, because they have enough physics and enough parameters available to accurately fit the measured I–V data. The real value of the conceptual model presented here is to provide a clear physical interpretation to the parameters in traditional models. Because compact models serve as a succinct description of our conceptual understanding of devices, one that device physicists use to interpret experiments and develop new technologies, it is important that even if expressed in traditional form, compact models be presented in a way that accurately reflects the underlying physics.

References

[1] C. C. McAndrew, "Practical modeling for circuit simulation," *IEEE J. Solid-State Circuits*, vol. 33, no. 3, pp. 439–448, Mar. 1998.

[2] C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integr. Circuits Signal Process.*, vol. 8, pp. 83–114, Jul. 1995.

[3] M. Chan, K. Y. Hui, C. Hu, and P. K. Ko, "A robust and physical BSIM3 non-quasi-static transient and AC small signal model for circuit simulation," *IEEE Trans. Electron Devices*, vol. 45, no. 4, pp. 834–841, Apr. 1998.

[4] G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, *et al.*, "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," *IEEE Trans. Electron Devices*, vol. 53, no. 9, pp. 1979–1993, Sep. 2006.

[5] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[6] M. Miura-Mattausch, H. Ueno, M. Tanaka, H. Mattausch, S. Kamashiro, T. Yamashita, *et al.*, "HiSIM: A MOSFET model for circuit simulation connecting device performance with technology," in *Proc. IEDM*, Dec. 2002, pp. 109–112.

[7] Y. Tsividis and C. McAndrew, *Operation and Modeling of the MOS Transistor*. Oxford, U.K.: Oxford Univ. Press, 2011.

[8] H. K. J. Ihantola and J. L. Moll, "Design theory of a surface field-effect transistor," *Solid-State Electron.*, vol. 7, no. 6, pp. 423–430, 1964.

[9] H. C. Pao and C. T. Sah, "Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors (MOST)," *Solid-State Electron.*, vol. 9, no. 10, pp. 927–937, 1966.

[10] J. R. Brews, "A charge sheet model of the MOSFET," *Solid-State Electron.*, vol. 21, no. 2, pp. 345–355, 1987.

[11] B. J. Sheu, D. L. Scharfetter, P.-K. Ko, and M.-C. Jeng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," *IEEE J. Solid-State Circuits*, vol. 22, no. 4, pp. 558–556, Aug. 1987.

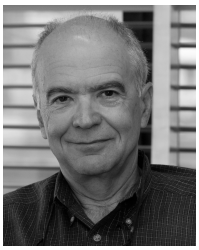
[12] J. D. Bude, "MOSFET modeling into the ballistic regime," in *Proc. Int. Conf. Simul. Semicond. Process. Devices*, 2000, pp. 23–26.

[13] D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo simulation of a 30 nm dual-gate MOSFET: How short can Si go?" in *IEDM Tech. Dig.*, Dec. 1992, pp. 553–556.

[14] M. R. Pinto, E. Sangiorgi, and J. Bude, "Silicon MOS transconductance scaling into the overshoot regime," *IEEE Electron Device Lett.*, vol. 14, no. 8, pp. 375–278, Aug. 1995.

[15] P. M. Solomon and S. E. Laux, "The ballistic FET: Design, capacitance, and speed limits," in *Proc. IEDM Tech. Dig.*, Dec. 2001, pp. 5.1.1–5.1.4.
[16] M. Lundstrom and Z. Ren, "Essential physics of carrier transport in nanoscale MOSFETs," *IEEE Trans. Electron Devices*, vol. 49, no. 1, pp. 133–141, Jan. 2002.
[17] L. Wei, O. Mysore, and D. A. Antoniadis, "Virtual-source based self-consistent current and charge FET models: From ballistic to drift-diffusion velocity-saturation operation," *IEEE Trans. Electron Devices*, vol. 59, no. 5, pp. 1263–1271, May 2012.
[18] A. Khakifirooz, O. M. Nayfeh, and D. A. Antoniadis, "A simple semiempirical short-channel MOSFET current–voltage model continuous across all regions of operation and employing only physical parameters," *IEEE Trans. Electron Devices*, vol. 56, no. 8, pp. 1674–1680, Aug. 2009.
[19] S. Datta, *Lessons from Nanoelectronics: A New Perspective on Transport*. Singapore: World Scientific, 2012.
[20] M. Lundstrom and C. Jeong, *Near-Equilibrium Transport: Fundamentals and Applications*. Singapore: World Scientific, 2013.
[21] E. O. Johnson, "The IGFET: A bipolar transistor in disguise," *RCA Rev.*, vol. 34, pp. 80–94, Mar. 1973.
[22] J.-H. Rhew, Z. Ren, and M. S. Lundstrom, "A numerical study of ballistic transport in a nanoscale MOSFET," *Solid-State Electron.*, vol. 46, no. 11, pp. 1899–1906, 2002.
[23] S. Eminente, D. Esseni, P. Palestri, C. Fiegna, L. Selmi, and E. Sangiorgi, "Enhanced ballisticity in nano-MOSFETs along the ITRS roadmap: A Monte Carlo study," in *IEDM Tech. Dig.*, 2004, pp. 609–612.
[24] E. Fuchs, P. Dollfus, G. Le Carval, S. Barraud, D. Villanueva, F. Salvetti, *et al.*, "A new backscattering model giving a description of the quasi-ballistic transport in nano-MOSFET," *IEEE Trans. Electron Devices*, vol. 52, no. 10, pp. 2280–2289, Oct. 2005.
[25] P. Palestri, D. Esseni, S. Eminente, C. Fiegna, E. Sangiorgi, and L. Selmi, "Understanding quasi-ballistic transport in nano-MOSFETs: Part I—Scattering in the channel and in the drain," *IEEE Trans. Electron Devices*, vol. 52, no. 12, pp. 2727–2735, Dec. 2005.
[26] H. Tsuchiya, K. Fujii, T. Mori, and T. Miyoshi, "A quantum-corrected Monte Carlo study on quasi-ballistic transport in nanoscale MOSFETs," *IEEE Trans. Electron Devices*, vol. 53, no. 12, pp. 2965–2971, Dec. 2006.
[27] R. Clerc, P. Palestri, and L. Selmi, "On the physical understanding of the kT-Layer concept in quasi-ballistic regime of transport in nanoscale devices," *IEEE Trans. Electron Devices*, vol. 53, no. 7, pp. 1634–1640, Jul. 2006.
[28] L. Lucci, P. Palestri, D. Esseni, L. Bergagnini, and L. Selmi, "Multisubband Monte Carlo study of transport, quantization, and electron-gas degeneration in ultrathin SOI n-MOSFETs," *IEEE Trans. Electron Devices*, vol. 54, no. 5, pp. 1156–1164, May 2007.
[29] J. Lusakowski, M. J. M. Martinez, R. Rengel, T. Gonzalez, R. Tauk, Y. M. Meziani, *et al.*, "Quasiballistic transport in nanometer Si metal-oxide-semiconductor field-effect transistors: Experimental and Monte Carlo analysis," *J. Appl. Phys.*, vol. 101, no. 11, pp. 114511-1–114511-6, 2007.
[30] R. Kim and M. S. Lundstrom, "Physics of carrier backscattering in one- and two-dimensional nanotransistors," *IEEE Trans. Electron Devices*, vol. 56, no. 1, pp. 132–139, Jan. 2009.
[31] T. Tsutsumi and K. Tomizawa, "Analysis of backscattering phenomena from drain region in silicon decanano diode," *Jpn. J. Appl. Phys.*, vol. 45, pp. 6786–6789, Sep. 2006.
[32] M. V. Fischetti, T. P. O'Regan, N. Sudarshan, C. Sachs, S. Jin, J. Kim, *et al.*, "Theoretical study of some physical aspects of electronic transport in n-MOSFETs at the 10-nm gate-length," *IEEE Trans. Electron Devices*, vol. 54, no. 9, pp. 2116–2136, Sep. 2007.
[33] M. V. Fischetti, L. Wang, B. Yu, C. Sachs, P. M. Asbeck, Y. Taur, *et al.*, "Simulation of electron transport in high-mobility MOSFETs: Density of states bottleneck and source starvation," in *Proc. IEEE IEDM*, Dec. 2007, pp. 109–112.
[34] M. V. Fischetti, S. Jin, T.-W. Tang, P. Asbeck, Y. Taur, S. E. Laux, *et al.*, "Scaling MOSFETs to 10 nm: Coulomb effects, source starvation, and virtual source model," *J. Comput. Electron.*, vol. 8, no. 2, pp. 60–77, 2009.
[35] D. H. Kim, J. A. del Alamo, D. A. Antoniadis, and B. Brar, "Extraction of virtual-source injection velocity in sub-100 nm III–V HFETs," in *Proc. IEEE IEDM*, Dec. 2009, pp. 861–864.
[36] K. Natori, "Ballistic metal-oxide-semiconductor field effect transistor," *J. Appl. Phys.*, vol. 76, no. 8, pp. 4879–4890, 1994.
[37] A. Rahman, J. Guo, S. Datta, and M. Lundstrom, "Theory of ballistic nanotransistors," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1853–1864, Sep. 2003.
[38] M. S. Lundstrom, *Fundamentals of Carrier Transport*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[39] M. Lundstrom and J. Guo, *Nanoscale Transistors: Physics, Modeling, and Simulation*. New York, NY, USA: Springer-Verlag, 2006.
[40] J. Wang and M. S. Lundstrom, "Ballistic transport in high electron mobility transistors," *IEEE Trans. Electron Devices*, vol. 50, no. 7, pp. 1604–1609, Jul. 2003.
[41] N. Neophytou, T. Rakshit, and M. S. Lundstrom, "Performance analysis of 60 nm gate-length III–V InGaAs HEMTs: Simulations versus experiments," *IEEE Trans. Electron Devices*, vol. 56, no. 7, pp. 1377–1387, Jul. 2009.
[42] F. Assad, Z. Ren, D. Vasileska, S. Datta, and M. S. Lundstrom, "On the performance limits of silicon MOSFETS: A theoretical study," *IEEE Trans. Electron Devices*, vol. 47, no. 1, pp. 232–240, Jan. 2000.
[43] A. Lochtefeld and D. A. Antoniadis, "On experimental determination of carrier velocity in deeply scaled NMOS: How close to the thermal limit?" *IEEE Electron Device Lett.*, vol. 22, no. 2, pp. 95–97, Feb. 2001.
[44] C. Jeong, D. A. Antoniadis, and M. Lundstrom, "On backscattering and mobility in nanoscale silicon MOSFETs," *IEEE Trans. Electron Devices*, vol. 56, no. 11, pp. 2762–2769, Nov. 2009.
[45] M.-J. Chen, H.-T. Huang, Y.-C. Chou, R.-T. Chen, Y.-T. Tseng, P.-N. Chen, *et al.*, "Separation of channel backscattering coefficients in nanoscale MOSFETs," *IEEE Trans. Electron Dev.*, vol. 51, no. 9, pp. 1409–1415, Sep. 2004.
[46] M. S. Lundstrom, "Elementary scattering theory of the Si MOSFET," *IEEE Electron Device Lett.*, vol. 18, no. 7, pp. 361–363, Jul. 1997.
[47] M. S. Shur, "Low ballistic mobility in submicron HEMTs," *IEEE Electron Device Lett.*, vol. 23, no. 9, pp. 511–513, Sep. 2002.
[48] A. Majumdar and D. A. Antoniadis, "Analysis of carrier transport in short-channel MOSFETs," submitted for publication, 2013.
[49] G. Mugnaini and G. Iannaccone, "Physics-based compact model of nanoscale MOSFETs—Part I: Transition from drift-diffusion to ballistic transport," *IEEE Trans. Electron Devices*, vol. 52, no. 8, pp. 1795–1801, Aug. 2005.
[50] G. Mugnaini and G. Iannaccone, "Physics-based compact model of nanoscale MOSFETs—Part II: Effects of degeneracy on transport," *IEEE Trans. Electron Devices*, vol. 52, no. 8, pp. 1802–1806, Aug. 2005.
[51] Y. Liu, M. Luisier, A. Majumdar, D. Antoniadis, and M. S. Lundstrom, "On the interpretation of ballistic injection velocity in deeply scaled MOSFETs," *IEEE Trans. Electron Devices*, vol. 59, no. 4, pp. 994–1001, Apr. 2012.
[52] J. Guo, S. Datta, M. Lundstrom, M. Brink, P. McEuen, A. Javey, *et al.*, "Assessment of silicon MOS and carbon nanotube FET performance limits using a general theory of ballistic transistors," in *Proc. IEDM*, Dec. 2002, pp. 711–714.
[53] R. Venugopal, S. Goasguen, S. Datta, and M. S. Lundstrom, "Quantum mechanical analysis of channel access geometry and series resistance in nanoscale transistors," *J. Appl. Phys.*, vol. 95, no. 1, pp. 292–305, 2004.
[54] U. Radhakrishna, L. Wei, D.-S. Lee, T. Palacios, and D. A. Antoniadis, "Physics-based GaN HEMT transport and charge model: Experimental verification and performance projection," in *Proc. IEEE IEDM*, Dec. 2012, pp. 319–322.
[55] C. C. McAndrew, private communication, May 2013.
[56] M. Lundstrom. (2012). *Nanoscale Transistors* [Online]. Available: https://nanohub.org/groups/u-fall2012-lundstrom01
[57] D. Antoniadis and S. Rakheja. (2013). *MVS 1.0.0 Nanotransistor Model (Silicon)* [Online]. Available: https://nanohub.org/resources/19223

Mark S. Lundstrom (S'72–M'74–SM'80–F'94) is the Don and Carol Scifres Professor of electrical engineering with Purdue University, West Lafayette, IN, USA, where he works on electronic devices. He is a member of the National Academy of Engineering.



Dimitri A. Antoniadis (M'79–SM'83–F'90) is the Ray and Maria Stata Professor of electrical engineering with the Massachusetts Institute of Technology, Cambridge, MA, USA, where he works on nanoelectronics. He is a member of the National Academy of Engineering.