New insights into carrier transport in n-MOSFETs

by A. Lochtefeld
I. J. Djomehri
G. Samudra
D. A. Antoniadis

This paper discusses recent experimental investigations of the relation between low-field effective mobility and effective injection velocity of electrons from the source into the channel, as manifested in current drive, of deeply scaled n-MOSFETs. It is first established that the effective velocity in electrostatically sound, "well-tempered" scaled devices, for example with drain-induced barrier lowering (DIBL) limited to 120 mV/V, is well below the theoretical fully ballistic injection velocity. This is consistent with the fact that, as the channel length is scaled and the longitudinal field increases, preservation of electrostatic integrity requires increasing transverse field, which leads to increased surface scattering and therefore decreased mobility. In addition, evidence is presented that the effective channel mobility in modern short-channel devices is further decreased, probably due to increased ionized dopant scattering in the heavily doped channel halos. Then a correlation range of 45–60% between effective injection velocity and low-field mobility is established experimentally in sub-50-nm-channel MOSFETs. All of these factors point to the possibility of increasing the performance of deeply scaled n-MOSFETs by pursuing enhanced channel-mobility device structures such as double-gate MOSFET, or materials such as strained Si on relaxed SiGe.

1. Introduction

The current drive capability of deeply scaled MOSFETs and, in particular, n-MOSFETs has been the subject of investigation since the late 1970s. First it was hypothesized that the effective carrier injection velocity from the source into the channel would reach the limit of the saturation velocity and remain there as longitudinal electric fields increased beyond the onset value for velocity saturation. However, theoretical work indicated that velocity overshoot can occur even in silicon [1], and indeed it is routinely seen in the high-field region near the drain in simulated devices using energy balance models or Monte Carlo. While it was understood that velocity overshoot near the drain would not help current drive, early experimental work [2, 3] claimed to observe velocity overshoot near the source, which of course would be beneficial. Velocity was extracted from the intrinsic saturation transconductance, g_{mi}, normalized to device width, W, and inversion capacitance per unit area, C'_{ox}, as g_{mi}/WC'_{ox}. A similar claim was made at the same time by Sai-Halasz et al. [4] by fitting the saturated velocity in a drift-diffusion simulator to match experimental device currents. For a period of time it appeared that with good channel doping engineering velocity overshoot near the source could become practical in deeply scaled devices,

0018-8646/02/$5.00 © 2002 IBM

but subsequent experimental work [5] demonstrated that there is a constraint between g_{mi}/WC'_{ox} (velocity) and drain-induced barrier lowering (DIBL). Indeed, velocity thus extracted could exceed the saturated velocity, but only at very high values of DIBL, where the devices would not be usable. That work also demonstrated the superiority of halos in the velocity vs. DIBL performance criterion, but for usable DIBL values, effective electron velocities have remained stubbornly well below the saturation velocity.

In this paper we revisit the issue of effective injection velocity and its relation to effective mobility using a combination of experiments and simulation. Strictly speaking, *effective mobility* is commonly obtained vs. effective transverse electric field from long-channel device measurements and is assumed to be constant along the channel. However, in modern short-channel devices with strong doping halos, mobility is not expected to be constant along the channel, so, by *effective mobility*, we mean the average mobility in the channel. An additional difficulty in extending the mobility concept to very-short-channel devices comes from the fact that the halo doping dimensions, and indeed the channel length itself, are only moderately greater than the electron mean free path. Nevertheless, since it is found that at low longitudinal fields, irrespective of doping scheme or channel length, current and thus carrier velocity are proportional to field, an effective mobility can always be defined. We interpret this effective mobility as a quantity proportional to the average carrier scattering rate. It is shown here that effective mobility (at low longitudinal field) and velocity (at high field) are correlated and are both degraded as bulk n-MOSFETs are scaled down, probably because of increased surface and ionized impurity scattering. Alternative device structures that may overcome this limitation are then discussed briefly.

2. Effective channel-injection velocity in sub-100-nm n-MOSFETs

Continued success in scaling bulk MOSFETs has brought increasing focus on fundamental device performance limits. The ultimate limit to performance is thought to be the thermal injection velocity v_θ ($1.2–2 \times 10^7$ cm/s) from the source accumulation layer into the channel [6, 7]. By applying the formalism of 1-flux scattering theory [6], the limit can be stated as

$$I_{on}/W = v_\theta Q_i(x_0)T/(2 - T), \tag{1}$$

where I_{on} is the saturated drain current and $Q_i(x_0)$ is the areal inversion layer density at the conduction-band peak at $x = x_0$ (with $V_{gs} = V_{ds} = V_{dd}$) at the source side of the channel. x denotes longitudinal channel position. The *effective* channel-injection carrier velocity at x_0 is

$$v_{eff} = v_\theta T/(2 - T). \tag{2}$$

T is the transmission coefficient at x_0; $T = 1$ (and $v_{eff} = v_\theta$) represents fully ballistic transport (i.e., no backscattering from the channel back to the source). As a measure of how close to the thermal limit a device operates, it is conceptually useful to define a thermal or ballistic efficiency, β:

$$\beta \equiv v_{eff}/v_\theta = T/(2 - T). \tag{3}$$

To estimate T and β, v_{eff} must be determined experimentally (v_θ can be estimated theoretically [7]) as close to the conduction-band peak as possible, if the goal is to assess how near to the ballistic limit a modern MOSFET operates. The answer to this question has important ramifications: Large β for a modern "standard" MOSFET would suggest that only minor drive-current benefit could be expected from continued scaling, or from technology alternatives for mobility improvement (e.g., strained Si or undoped thin-film SOI), as we later discuss. In the following sections we discuss several experimental methods for estimating v_{eff}.

Carrier velocity from saturated transconductance

Effective carrier velocity can be measured from extrinsic or intrinsic saturated transconductance g_m and g_{mi} [3]:

$$v_{gm} = g_m/WC'_{ox}, \tag{4}$$

$$v_{gmi} = g_{mi}/WC'_{ox}, \tag{5}$$

where C'_{ox} is the gate-oxide capacitance per unit area in inversion, g_{mi} is the saturated transconductance corrected for source/drain parasitic resistance (R_{sd}) as in [8], and W is the width of the device. v_{gmi}, corrected for R_{sd}, is a more accurate reflection of real channel carrier velocity than v_{gm}.

Carrier velocity from drain current and long-device CV

Conceptually, a more straightforward way to extract v_{eff} is to directly measure $I_{on}/WQ_i(x_0)$. Determining $Q_i(x_0)$ in deeply scaled devices is problematic because of uncertainties in channel length, large (relative) overlap and fringing capacitances, and nonuniform charge distribution along the channel. However, in strong inversion and in the gradual channel approximation, $Q_i(x_0)$ in a short channel should correspond closely to the long-channel inversion layer charge $\int_0^{V_{gs}} C'_{gsd}|_{V_{ds}=0}$, where C'_{gsd} is the gate-to-source/drain (tied) capacitance, normalized to unit area. Choosing a device with a sufficiently long channel renders the fringing component of C'_{gsd} negligible. Accordingly, we let

$$Q_i(x_0)_{(\text{short-chan.})} = \int_0^{V_{gs^*}} C'_{gsd}|_{V_{ds}=0(\text{long-chan.})}, \quad (6)$$

where $V_{gs}^* = V_{gs} + \Delta V_{gs}$, with ΔV_{gs} accounting for differences between long- and short-channel devices. The most important component in ΔV_{gs} is ΔV_t due to drain-induced barrier lowering (DIBL) and threshold-voltage rolloff. The expression for effective velocity as extracted from I_{on} becomes

$$v_{id} = I_{on}/WQ_i(x_0)_{(\text{short-chan.})} = I_{on}\Big/\left[W\int_0^{(V_{gs}+\Delta V_t)} C'_{gsd(\text{long-chan.})}\right]. \quad (7)$$

A second component in ΔV_{gs} is due to voltage drop on the source resistance, $I_{on}R_s$. Adding this correction to the upper integration limit in Equation (7) gives the expression used for "intrinsic" effective velocity, v_{idi}:

$$v_{idi} = I_{on}\Big/\left[W\int_0^{(V_{gs}+\Delta V_t-I_{on}R_s)} C'_{gsd(\text{long-chan.})}\right]. \quad (8)$$

Carrier velocity from drain current and short-device CV

A carrier velocity extraction technique was presented in [9] which we denote v_{id2}:

$$v_{id2} = I_{on}/WQ_i = I_{on}\Big/\left[W\int_{V_t}^{V_{gs}} C'_{gs(V_{ds}=V_{dd})}\right]. \quad (9)$$

C'_{gs} is the capacitance (per unit area) measured from the short device. Despite important advantages, this technique is difficult to apply to short-channel devices because of a significant fringing capacitance correction and the need to know L_{eff} accurately in order to normalize C'_{gs}. Even when accurately applied, this technique gives an average carrier velocity in the channel and not the velocity near x_0, as discussed below.

A simulated MOSFET was used to compare the different velocity-extraction methods. Each method was simulated exactly, and each extracted velocity was marked on the simulated velocity plot as x location, thus identifying the channel location for each technique. These are shown in **Figure 1** for two different simulation models, drift-diffusion (DD) and energy balance (EB). The simulation results show clearly that the extraction methods developed in this work (v_{id}, v_{idi}) give inversion-layer carrier velocities closer to x_0 than the methods from the literature (v_{gm}, v_{gmi}, v_{id2}). However, v_{id} and v_{idi} also correspond to points in the channel somewhat beyond x_0. We must interpret, then, the subsequent experimental results (based on v_{idi}) as putting a reasonable *upper bound* on v_{eff}, and therefore β, for the technologies investigated.

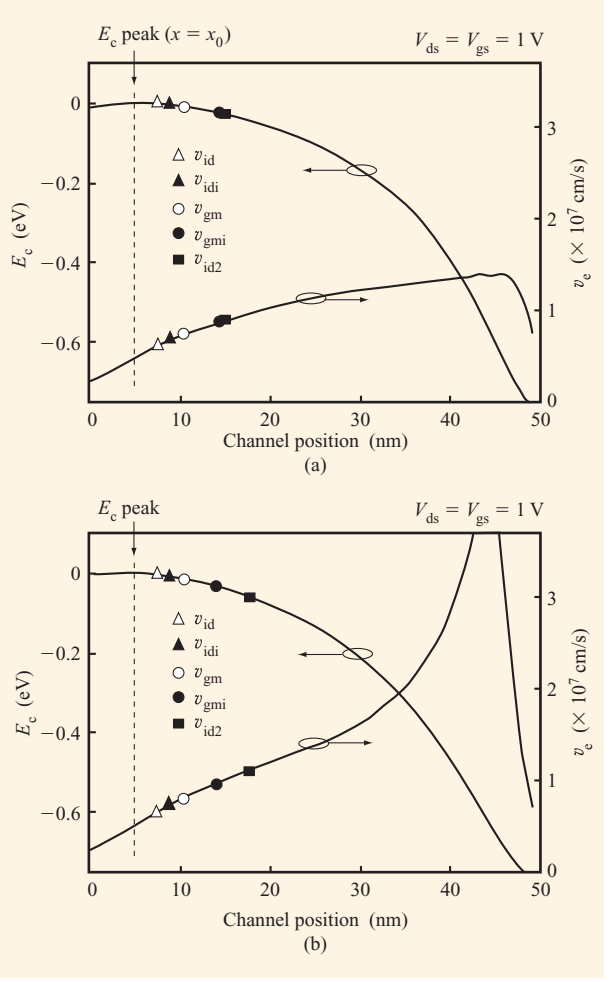

Figure 1

Simulation results: conduction-band and carrier velocity vs. position for a bulk "superhalo" device. Velocities at five marked points along the channel correspond to values given by different MOSFET carrier velocity extraction techniques. Solid symbols are corrected for source/drain series resistance; open symbols are uncorrected. Values for V_{th}, T_{ox}^{elec}, R_{sd}, and I_{on} approximately match those for Technology A discussed in the section on experimental results. DIBL here is 66 mV/V. (a) Drift-diffusion model: $v_{sat} = 1.3 \times 10^{-7}$ cm/s; (b) energy-balance model: $\tau_w = 0.1$ ps.

Experimental results

We measure v_{id} and v_{idi} (and compare with transconductance methods) for n-MOS devices from two advanced CMOS technologies, referred to in this work as technologies A and B (**Table 1**).[1] T_{ox}^{elec} was determined experimentally from $\varepsilon_{ox}/C'_{gsd}$ (with $V_{gs} = V_{dd}$) measured in a large ($L/W = 10\ \mu$m/10 μm) device. Source–drain parasitic series resistance values (R_{sd}) presented in Table 1 are estimated from inverse modeling [11, 12].

[1] These devices were obtained courtesy of industrial partners.

Table 1 Parameters for n-MOS technologies investigated.

Technology	T_{ox}^{elec} (nm)	V_t (V_{ds} = 50 mV, linear extrapolation, long chan.) (V)	*Nominal* V_{dd} (V)	R_{sd} (Ω-μm)	L_{eff} for DIBL <130 mV/V (nm)
A	2.4	0.3	1.0	190–220	~40
B	4.3	0.35	1.8	240–270	~65

Some uncertainty in the technique is reflected by the presentation of a range of values for each technology. The use of these ranges on series resistances introduces negligible error for the *relative* comparisons between measured velocity for technology A vs. B. Absolute error in v_{id} and v_{idi} corresponding to this uncertainty in R_{sd} is at most 1.0–1.5%.

To determine $Q_i(x_0)$ experimentally, we integrate C'_{gsd} obtained from a large (10-μm × 10-μm) device, and make adjustments according to Equation (6). C'_{ox} for v_{gm} and v_{gmi} was determined experimentally, from C'_{gsd} ($V_{gs} = V_{dd}$). The dependence of R_{sd} on gate bias is not taken into account because modest inaccuracies in R_{sd} do not significantly affect the results. Experimental results for technology A (**Figure 2**) [13] corroborate the simulation results, showing relative differences between v_{id}, v_{idi}, v_{gm}, and v_{gmi} of the same order. Similar results were found for the longer-channel technology B but with moderately less spread among the values. The fact that this difference is most pronounced at shorter channel lengths (either within one technology, or moving from B to A) suggests that with deeper scaling, effective velocity extraction via an $I_{on}/WQ_i(x_0)$ method such as v_{id} or v_{idi} is increasingly necessary.

Effective velocity in comparison to ballistic limit

The thermal injection velocity is a function of channel doping and inversion-layer density [7], increasing with both. Our estimates for v_θ are estimated from [7]. Using $v_\theta = 1.7 \times 10^7$ cm/s for technology A, and taking $v_{eff} = v_{idi}$ from Figure 2, we find that for DIBL = 100 mV/V, $\beta = v_{eff}/v_\theta = 0.39$, corresponding to an upper bound on T of 0.56. **Table 2** summarizes experimental results for technologies A and B, as well as for 25-nm (L_{eff}) Monte Carlo simulation results from [10], all at the same DIBL of 100 mV/V. It is important to compare velocities of different technologies at equal DIBL (regardless of measurement technique), because for a given technology carrier velocity increases as electrostatic integrity decreases [5]. The experimental results suggest that β is not increasing as we scale to shorter-channel-length generations. And, from the Monte Carlo results, it appears that with continued scaling (to 25 nm), bulk-Si n-MOS current drive would not be significantly above 40% of the thermally limited value. These results for technology B ($\beta = 0.47$) are consistent with reported results[2] [14].

In order to separate the effects of L_{eff} scaling from the corresponding changes in longitudinal electric field, the above experiments were repeated for different V_{dd} ($= V_{gs} = V_{ds}$). It is significant to note that, above $V_{dd} = 1.0$ V for technology A or 1.2 V for B, there is relatively little increase in carrier velocity with increasing drain bias. One possible explanation is the "tyranny of universal mobility," whereby electrostatic integrity requires increasing transverse electric field, and therefore reduced mobility, as the channel length is scaled and the longitudinal field increased. Another explanation is offered by the scattering theory approach to estimating MOSFET drain current [6]. In this view, for a MOSFET in strong inversion and with high drain bias, I_{on} is only weakly dependent upon longitudinal field and therefore drain bias [15].

3. Mobility, scaling, and the low-field mobility–velocity relationship

Although longitudinal electric fields in the channel of a modern MOSFET are far in excess of the E_{sat} value that leads to velocity saturation, theoretical [6, 16] and experimental work [17, 18], as well as our own work described later in this paper, suggests that carrier velocity at or near x_0 still depends strongly on μ_{eff} in the sub-100-nm regime. We note, however, that there is no universal agreement about this strong correlation of carrier velocity with μ_{eff}, e.g. [19].

In bulk-Si and SOI MOSFETs, the effective low longitudinal field mobility, μ_{eff} (typically extracted from long-channel devices), behaves according to a "universal" relationship depending only on E_{eff}, the effective or average transverse field seen by carriers in the inversion layer [20–22]:

[2] Mark Lundstrom, personal communication, March 2001.

Table 2 Ballistic efficiency β and transmission coefficient T. $\beta \equiv v_{eff}/v_{\theta} = T/(2 - T)$.

Properties	*25-nm Monte Carlo*	*Technology A*	*Technology B*
V_{eff} (cm/s)	$6.7–7.6 \times 10^6$	6.6×10^6	7.7×10^6
β	0.35–0.40	0.39	0.47
T	0.52–0.57	0.56	0.64

$$\mu_{eff} = \frac{\mu_0}{\left(1 + \left|\frac{E_{eff}}{E_0}\right|^{\nu}\right)}, \tag{10}$$

$$E_{eff} = \frac{(\eta Q_i + Q_b)}{\varepsilon_{si}}, \tag{11}$$

where μ_0, E_0, ν, and η are fitting parameters which depend on carrier type. $\nu = 1.6$ and $\eta = 0.5$ for electrons [21]. Q_i and Q_b are the inversion and channel depletion region charge areal densities, respectively. Typically for channel doping heavier than $2–3 \times 10^{18}$ cm^{-3}, Q_b becomes the dominant contribution to E_{eff}.

Considering a wide range of temperature and transverse field conditions, the dominant scattering mechanisms for carriers in MOSFET inversion layers are Coulomb, phonon, and surface-roughness scattering [23–25]; we can therefore approximate by Matthiesson's rule [26]:

$$\frac{1}{\mu} = \frac{1}{\mu_{coulomb}} + \frac{1}{\mu_{phonon}} + \frac{1}{\mu_{sr}}. \tag{12}$$

In modern MOSFETs at room temperature, the "universal" mobility behavior is thought to be dominated by surface roughness and $\mu_{sr} \ll \mu_{coulomb}$ [27]. For deeply scaled MOSFETs, this may not be the case, and this issue is examined here with the help of measurements.

Experimental determination of low-field mobility

For small $V_{ds} \ll V_{gs} - V_t$, it is well known that

$$I_d = \mu_{eff} C'_{ox} \frac{W}{L_{eff}} (V_{gs} - V_t) V_{ds}, \tag{13}$$

where L_{eff} is the effective channel length [28, 29]. As discussed in the Introduction, μ_{eff} for short-channel MOSFETs with strong halos should be considered (for small V_{ds}) as a proportionality ratio between electron velocity and longitudinal electric field, which at the long-channel limit becomes the well-defined lumped inversion charge mobility. Nevertheless, for either short or long devices, it is intimately related to carrier scattering rate, and as such it has meaning in either regime. Using the approximation $Q_i \approx C'_{ox}(V_{gs} - V_t)$, which is accurate in strong inversion, effective mobility can be determined according to

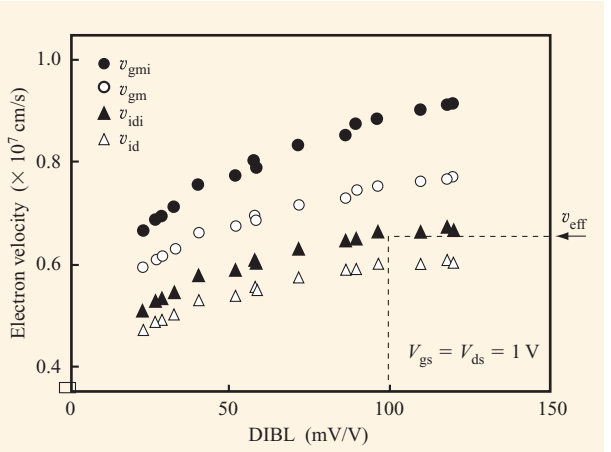

Figure 2

Experimental results: carrier velocity by four techniques vs. drain-induced barrier lowering (DIBL), for technology A. Solid symbols are corrected for source/drain series resistance; open symbols are uncorrected. The implicit variable in the DIBL axis is L_{gate}. Reprinted with permission from [13]; ©2001 IEEE.

$$\mu_{eff} = \frac{I_d L_{eff}}{V^*_{ds} W Q_i}, \tag{14}$$

where Q_i is the low-V_{ds} inversion charge areal density obtained experimentally and assumed to be uniform throughout the channel. V^*_{ds} $(= V_{ds} - I_d R_{sd})$ is the effective or intrinsic drain bias. For this investigation, both L_{eff} and R_{sd} are extracted from short-channel devices via inverse modeling [11, 12]. This, together with the determination of Q_i from Equation (6), allows Equation (14) to be used to determine mobility in short-channel devices.

We apply Equation (14) to measure the μ_{eff} vs. L_{eff} relationship for technology A. The long-channel devices obey the universal mobility relations quite well. To explore mobility behavior in short-channel devices for several values of constant inversion charge Q_i, we measure μ_{eff} via Equation (14) at three gate overdrives $V_{od} = V_{gs} - V_t$ (where V_t is the linearly extrapolated value at $V_{ds} = 50$ mV). This is shown in **Figure 3**. Effective mobility at short channels appears to be independent of gate bias and inversion-charge density. This points toward strong Coulomb scattering at short channel lengths.

The mobility measured for two constant E_{eff} values is shown in **Figure 4** [30]. This clearly demonstrates the disappearance of universal mobility behavior at short channels. This behavior, as well as the trend of μ_{eff}

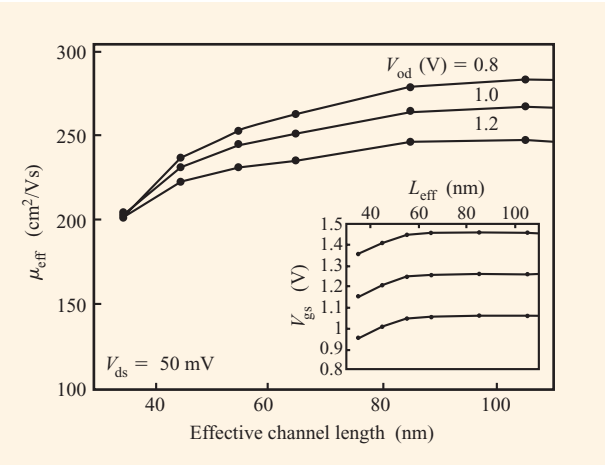

Figure 3

Measured μ_{eff} vs. L_{eff} for different gate overdrive (technology A). Applied V_{gs} is indicated in inset.

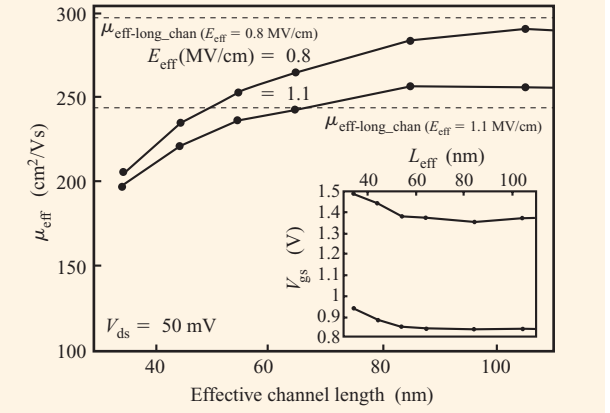

Figure 4

Measured μ_{eff} vs. L_{eff} for different E_{eff} (technology A). Applied V_{gs} is indicated in inset. Reprinted with permission from [30].

degradation (up to about 30% for the lower E_{eff} value), is interpreted as evidence that $\mu_{coulomb} < \mu_{sr}$. This is plausible because of the heavy source and drain halo dopings in the channels of these devices which merge for very short gate lengths. Therefore, device architectures that may allow undoped channels (see later) would be beneficial. On the other hand, the electron mobility may be suffering from long-range Coulomb interactions with electrons in the heavily doped source/drain and gate regions [19]. If that is the case, undoped channels would not yield a significant benefit.

The results of μ_{eff} versus channel length presented in this section should be considered preliminary for two reasons: sensitivity of experimental μ_{eff} to error in R_{sd} and L_{eff} estimations, and irregularities in asymptotic behavior of the results at longer channel lengths.

Electron velocity dependence on mobility in deep-sub-100-nm bulk n-MOS

The relation of low-field mobility to the performance of deep-sub-100-nm MOSFETs is still controversial. Here we investigate experimentally, for electrons in short (45-nm) n-MOS devices, the relation between mobility at low longitudinal electric fields (μ_{eff}) and velocity in the MOSFET saturation regime (v_{eff}), where peak longitudinal fields in the channel are high.

We investigate n-MOS transistors in the 1-V CMOS technology A, with L_{eff} for electrostatically sound devices (DIBL $\leqq$ 120 mV/V) down to ~45 nm. Using a four-point bending apparatus, compressive and tensile uniaxial stress parallel to the direction of electron transport is applied to a silicon strip containing several processed dies. Surface strains of up to 0.12% are achieved. This method allows electrical characterization of the same set of devices with and without strain, reducing sources of experimental error. Fractional change in low-field effective mobility ($\delta\mu_{eff} \equiv \Delta\mu_{eff}/\mu_{eff}$) corresponding to the induced strain is measured in long (10-μm) and short (45-nm) devices, as shown in **Figure 5**, using the technique described earlier. Some of the problems associated with mobility measurement in short devices—difficulty in accurately determining L_{eff} and Q_i [Equation (14)]—are not a significant problem for this experiment, because only the ratio of strained to unstrained mobility is required. When $\mu_{eff\text{-}strained}/\mu_{eff\text{-}unstrained}$ is measured, the uncertainties in L_{eff} and Q_i cancel out. This "cancellation of uncertainties" does not apply to the drain-bias term. Determination of $\delta\mu_{eff}$ in the long-channel device is not significantly affected by this. However, the sensitivity to R_{sd} is evident in the greater scatter among data points for $\delta\mu_{eff\text{-}short}$: For each measurement at a new strain value, the probe-tip-to-pad contact resistance varies, slightly changing the total source/drain series resistance. This is also why the straight-line fit to the data does not pass through (0, 0) in Figure 5 for the short devices. The difference between $\delta\mu_{eff\text{-}long}$ and $\delta\mu_{eff\text{-}short}$—a 40% reduction of dependence on strain for μ_{eff} in the short devices—may be indicative of a transition in the dominant scattering mechanism with device scaling, as discussed earlier, and is not unlike the Si piezoresistance coefficient reduction with increased doping [31].

Effective velocity versus mobility

Experimentally extracted v_{gmi} and v_{idi} are plotted against mobility shift for the short device in technology A, as shown in **Figure 6** [32]. We denote the ratio $\delta v_e/\delta\mu_{eff}$ as $R_{v\mu}$, and interpret it as a measure of the dependence of source-end electron velocity in the MOSFET saturation regime on low-field mobility. From Figure 6, $R_{v\mu} = 0.46$–0.48. If this is calculated with long-channel mobility, $R_{v\mu} = 0.3$, which is much lower. Clearly, for understanding transport in deep-sub-100-nm MOS devices, it is not valid to infer short-device mobility behavior from measurements of long devices from the same technology.

Earlier theoretical work with energy transport models [16] relating mobility to velocity agrees approximately, at the 50-nm-channel-length node, with our experimental results ($R_{v\mu} \approx 0.5$). We also performed 2D device simulations to support the measured results. Energy-balance (EB) modeling of a realistic simulated superhalo n-MOSFET with 2D doping profiles carefully designed to match measured subthreshold characteristics (DIBL, subthreshold slope, I_{off}) of the short (45-nm) devices with assumed change in the energy relaxation time $\delta\tau_w = \delta\mu_{eff}$ (where $\delta\tau_w \equiv \Delta\tau_w/\tau_w$), results in $R_{v\mu} = 0.55$, reasonably close to the experimental value. This assumption of $\delta\tau_w \approx \delta\mu_{eff}$ has previously been used to successfully model Si/SiGe strained-Si MOSFETs [18].

If the piezoresistance effect in the source and drain resistance, R_{sd}, is accounted for with the help of a resistance test structure and inverse modeling, we find $R_{v\mu}$ is increased from 0.47 to 0.59 [32]. It is therefore reasonable to expect an $R_{v\mu}$ within the bounds of 0.45 and 0.60, which clearly indicates that increasing the low-field effective channel mobility continues to be beneficial even at $L_{eff} = 45$ nm.

4. Increased effective velocity in deeply scaled n-MOSFETs

From our observations, it appears that performance can improve if one can free the devices from the tyranny of Si universal mobility either by reducing E_{eff} while maintaining electrostatic integrity, or by shifting the whole mobility vs. E_{eff} curve up, as for example by use of biaxial strain (e.g., [18]). Since increased mobility leads to an increase in channel carrier velocity, the performance and ballistic efficiency will improve. Because a variety of options are being explored, and there are detailed papers on those in this issue, we discuss these options only briefly.

Step-doped bulk MOSFET

A hypothetical alternative to the nonuniformly doped channel is the perfect step-doped bulk MOSFET (also known as the ground-plane MOSFET [33, 34]), where channel doping, $N_b = 0$, down to depth T_{step}, and is arbitrarily high beyond. Maintaining other characteristics the same, E_{eff} is reduced in this structure, and hence mobility is enhanced over that of a uniformly doped bulk MOSFET, as shown in **Figure 7**. Note that in any doping scenario the E_{eff} near the source, where the mobility has its highest impact on velocity, is much reduced relative to mid-channel E_{eff} in a long device because of bulk charge-sharing with the source and also with the drain if the channel is short enough. Neglecting this 2D charge-

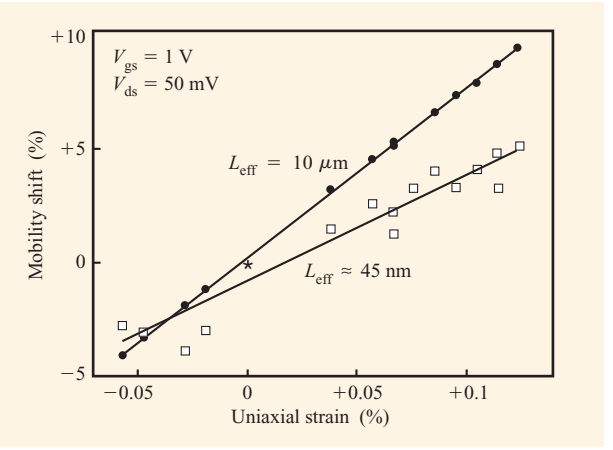

Figure 5

Experimental results: normalized mobility shift in long and short ($L_{eff} = 10\,\mu$m, ~45 nm) devices vs. uniaxial strain (technology A). Each point represents a fractional shift relative to unstrained value (which is represented by *).

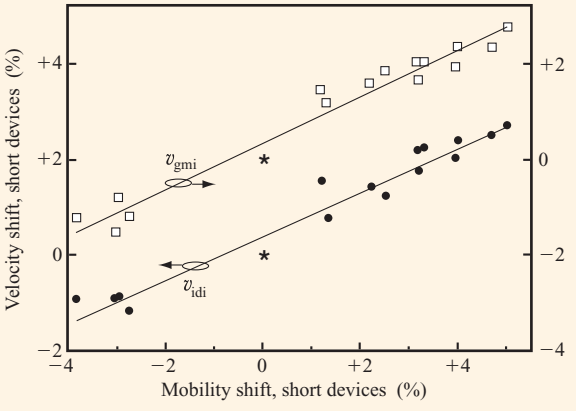

Figure 6

Experimental results: δv_e extracted by two methods, vs. short-device $\delta\mu_{eff}$ (technology A). Gate and drain biases are as described for Figure 5. Adapted with permission from [32]; ©2001 IEEE.

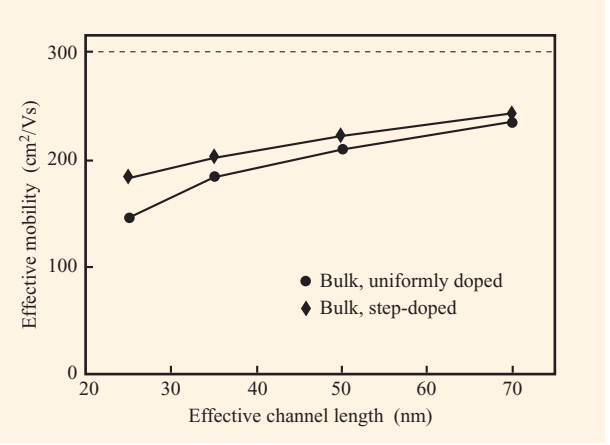

Figure 7

Calculated effective mobilities for two bulk n-MOSFET doping alternatives, correlated with scaling such that DIBL = 100 mV/V. In all cases the same mobility vs. E_{eff} was used. Dashed line corresponds to the hypothetical case of $Q_b = 0$ and $\Phi_{bi} = 0$, applicable only to SG and DG fully depleted SOI devices with mid-gap work-function gates.

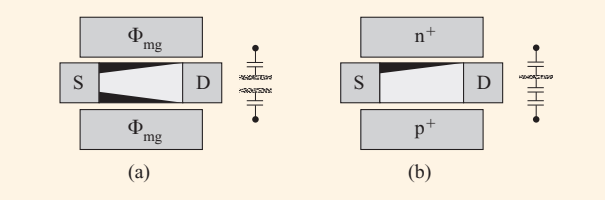

Figure 8

Alternative double-gate FDSOI structures: (a) DG–Φ_{mg}; (b) DG–n^+p^+.

sharing effect greatly overestimates the effect of channel doping on E_{eff} in short-channel devices. Therefore, the proper effective mobility can only be calculated from full 2D simulations. However, μ_{eff} is still much lower than in the case where there is no contribution to E_{eff} from the bulk charge. This limit is indicated by the dashed line without symbols in Figure 7. While there is no combination of channel and halo doping with gate work function that can reach this limit in scaled bulk devices, a potential alternative is the double-gate (DG) fully depleted SOI MOSFET, as discussed next.

Single- and double-gate fully depleted SOI MOSFET

The two alternatives for double-gate (DG) fully depleted SOI (FDSOI) design are illustrated schematically in **Figure 8**. Symmetrical DG–Φ_{mg} (mg denotes mid-gap gate work function) has two inversion layers, while DG–n^+p^+ and single-gate (SG) FDSOI have one. In fully depleted SOI devices (Figure 8), short-channel effects are suppressed by limiting silicon and oxide film thicknesses. The deleterious effect of drain bias on source-side channel potential is limited by device geometry so that the requirement of the gradual channel is relaxed; FDSOI devices thus do not suffer as much from the "tyranny of universal mobility." If alternative gate material processes can be developed such that gate work functions alone set an acceptable threshold voltage, Q_b can be essentially zero, and the μ_{eff}–E_{eff} range of operation is decoupled from device scaling. Figure 7 includes E_{eff} and μ_{eff} corresponding to the limits $Q_b = 0$ and $\Phi_{bi} = 0$, illustrating this potential benefit of reduced E_{eff} in FDSOI SG and DG MOSFETs as compared to bulk.

One might question the validity of using the bulk-Si μ_{eff}–E_{eff} curve for ultrathin-film SOI and DG–SOI devices. Provided the film is thicker than about 8 nm, as is the case for the devices considered here, it can be easily shown by the solution of coupled Poisson–Schrödinger equations that the presence of the second interface hardly modifies the inversion charge distribution of the first interface. Hence, the relationship between E_{eff} and the charge centroid distance from the interface is nearly the same as in bulk Si. In addition, recent experimental work with long-channel SOI devices with various Si film thicknesses [35] has shown that the bulk-Si universal mobility holds for film thicknesses down to about 9 nm.

For SG–Φ_{mg} with $Q_b \ll Q_i$, $E_{eff} = Q_i/2\varepsilon_{si}$. For DG–$\Phi_{mg}$, E_{eff} is further reduced by half, if Q_i is interpreted as the sum of inversion-layer charge at both gate interfaces: $E_{eff} = Q_i/4\varepsilon_{si}$. This indicates an important difference between transport in DG–Φ_{mg} and DG–n^+p^+ devices. Since the latter has a single inversion layer (at the n^+ interface for n-MOS), E_{eff} will be twice as high when compared with a DG–Φ_{mg} device with the same total Q_i, resulting in a reduced μ_{eff}. Investigations have shown that DG FDSOI MOSFETs may ultimately scale to ~10 nm channel length [36]. For double-gate devices, our study indicates that hypothetical mid-gap top and bottom gates are superior to n^+/p^+ polysilicon gates in terms of both scalability and drive current.

The mobility-enhancement ranges achievable through these modified structures are summarized in **Figure 9** at two different channel lengths of 50 and 25 nm. The advantages of the double-gate structure are very clear.

Biaxially strained Si-channel MOSFETs

The universal mobility limitations also can be overcome with the use of strained silicon on relaxed silicon–germanium by substantially shifting up the universal mobility curve. Biaxial strain raises electron mobility in n-MOSFETs much above the universal Si MOS curve.

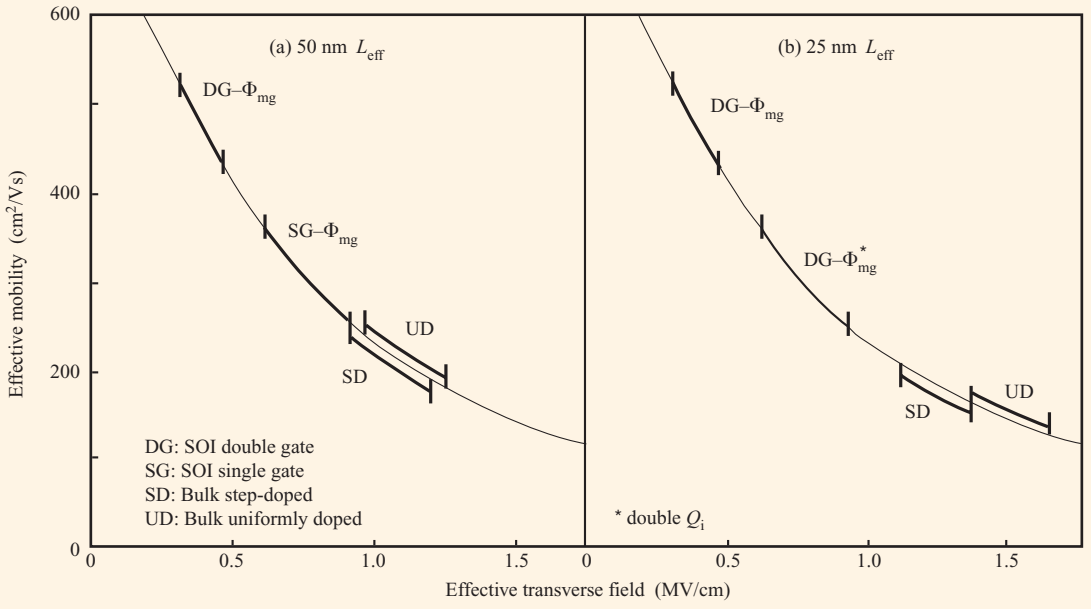

Figure 9

Universal effective n-MOSFET mobility (for low longitudinal field) from Equation (10), with regions of operation delineated for four different bulk and FDSOI device architectures. For each device architecture, the range of E_{eff} corresponds to $0.8 \times 10^{13} < Q_i/q < 1.2 \times 10^{13}$ cm^{-2}. In the case of DG–Φ_{mg}, Q_i is the sum of inversion charge in both inversion layers, except for the special case noted in (b). Threshold voltage is assumed 0.2 V at high V_{ds} for all device types.

A mobility-enhancement ratio of ~1.75 at high transverse fields has been reported [18].

5. Conclusion

In summary, we have demonstrated a technique for measuring effective carrier velocity near the source side of the MOSFET channel which is more appropriate than existing techniques for the purpose of determining how close modern technologies are to the thermal (ballistic) limit. With this we have shown that a deeply scaled (L_{eff} < 50 nm) 1-V n-MOS technology operates, at most, at 40% of the limiting thermal velocity.

We have shown that mobility in the shortest n-MOSFETs from a deep-sub-100-nm technology does not behave according to a traditional universal relationship with E_{eff}. We interpret this as evidence that Coulomb scattering (perhaps from ionized channel impurities, but possibly from electrons in the source, drain, and gate regions) is limiting the mobility. Also, by corroborating measured velocity and mobility dependence on strain, we have demonstrated experimentally the importance of low-field effective inversion-layer mobility in deep-sub-100-nm bulk n-MOS in increasing the effective velocity. Thus, the ability of SG- and DG-FDSOI to maintain high mobilities with deep scaling over bulk becomes a very significant benefit. Hence, there are a variety of alternatives available to improve on ballistic efficiency by increasing effective mobility and, hence, effective channel velocity.

Acknowledgment

This work was supported by the DARPA AME Program and SRC grants.

References

1. J. G. Ruch, "Electron Dynamics in Short-Channel Field-Effect Transistors," *IEEE Trans. Electron Devices* **ED-19,** 652 (1972).
2. S. Chou, D. Antoniadis, and H. Smith, "Observation of Electron Velocity Overshoot in Sub-100-nm-Channel MOSFETs in Si," *IEEE Electron Device Lett.* **EDL-6,** 665 (1985).
3. G. Shahidi, D. Antoniadis, and H. Smith, "Electron Velocity Overshoot at Room and Liquid Nitrogen Temperatures in Silicon Inversion Layers," *IEEE Electron Device Lett.* **8,** 94 (1988).
4. G. A. Sai-Halasz, M. F. Wordeman, D. P. Kern, S. Rishton, and E. Gamin, "High Transconductance and Velocity Overshoot in nMOS Devices at 0.1 μm Gate-Length Level," *IEEE Electron Device Lett.* **8,** 464 (1988).
5. H. Hu, J. Jacobs, L. Su, and D. Antoniadis, "A Study of Deep-Submicron MOSFET Scaling Based on Experiment and Simulation," *IEEE Trans. Electron Devices* **42,** 669 (1995).

6. M. Lundstrom, "Scattering Theory of the Short Channel MOSFET," *IEDM Tech. Digest*, p. 387 (1996).
7. F. Assad, Z. Ren, S. Datta, M. Lundstrom, and P. Bendix, "Performance Limits of Silicon MOSFET's," *IEDM Tech. Digest*, p. 547 (1999).
8. S. Chou and D. Antoniadis, "Relationship Between Measured and Intrinsic Transconductance of FETs," *IEEE Trans. Electron Devices* **ED-34,** 448 (1987).
9. T. Mizuno and R. Ohba, "Experimental Study of Nonstationary Electron Transport in Sub-0.1 μm Metal-Oxide-Silicon Devices: Velocity Overshoot and Its Degradation Mechanism," *J. Appl. Phys.* **82,** 5235 (1997).
10. Y. Taur, C. Wann, and D. Frank, "25 nm CMOS Design Considerations," *IEDM Tech. Digest*, p. 789 (1998).
11. Z. Lee, "A New Inverse-Modeling-Based Technique for Sub-100-nm MOSFET Characterization," Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, November 1998.
12. Z. Lee, M. McIlrath, and D. Antoniadis, "Two-Dimensional Doping Profile Characterization of MOSFETs by Inverse Modeling Using I–V Characteristics in the Subthreshold Region," *IEEE Trans. Electron Devices* **46,** 1640 (1999).
13. A. Lochtefeld and D. Antoniadis, "On Experimental Determination of Carrier Velocity in Deeply Scaled NMOS: How Close to the Thermal Limit?," *IEEE Electron Device Lett.* **22,** 96-97 (2001).
14. F. Assad, Z. Ren, D. Vasileska, S. Datta, and M. Lundstrom, "Modeling On-Currents for n-MOSFETs: Ultimate Limits vs. the NTRS," *Proceedings of the 1999 International Conference on Modeling and Simulation of Microsystems*, San Juan, Puerto Rico, 1999, p. 388.
15. S. Datta, F. Assad, and M. Lundstrom, "The Silicon MOSFET from a Transmission Viewpoint," *Superlatt. & Microstruct.* **23,** 771 (1998).
16. M. Pinto, E. Sangiorgi, and J. Bude, "Silicon MOS Transconductance Scaling into the Overshoot Regime," *IEEE Electron Device Lett.* **13,** 375 (1993).
17. T. Mizuno and R. Ohba, "Physical Limitations and Design for Sub-0.1μm MOS Devices: Carrier Velocity Overshoot and Performance Fluctuations," *Electron. & Commun. Jpn*., Part 2, **81,** 18 (1998).
18. K. Rim, J. Hoyt, and J. Gibbons, "Fabrication and Analysis of Deep Submicron Strained-Si n-MOSFETs," *IEEE Trans. Electron Devices* **47,** 1406 (2000).
19. M. Fischetti and S. Laux, "Performance Degradation of Small Silicon Devices Caused by Long-Range Coulomb Interactions," *Appl. Phys. Lett.* **76,** 2277 (2000).
20. A. Sabnis and J. Clemens, "Characterization of the Electron Mobility in the Inverted ⟨100⟩ Si Surface," *IEDM Tech. Digest*, p. 18 (1979).
21. M. Liang, J. Choi, P. Ko, and C. Hu, "Inversion-Layer Capacitance and Mobility of Very Thin Gate-Oxide MOSFET's," *IEEE Trans. Electron Devices* **ED-33,** 409 (1986).
22. M. Sherony, L. Su, J. Chung, and D. Antoniadis, "SOI MOSFET Effective Channel Mobility," *IEEE Trans. Electron Devices* **41,** 276 (1994).
23. C. Sah, T. Ning, and L. Tschopp, "The Scattering of Electrons by Surface Oxide Charges and by the Lattice Vibrations at the Si–SiO_2 Interface," *Surf. Sci.* **32,** 561 (1972).
24. S. Sun and J. Plummer, "Electron Mobility in Inversion and Accumulation Layers on Thermally Oxidized Silicon Surfaces," *IEEE Trans. Electron Devices* **ED-27,** 1497 (1980).
25. T. Ando, A. Fowler, and F. Stern, "Electronic Properties of Two-Dimensional Systems," *Rev. Mod. Phys.* **54,** 437 (1982).
26. D. Jeon and D. Burk, "MOSFET Electron Inversion Layer Mobilities—A Physically Based Semi-Empirical Model for a Wide Temperature Range," *IEEE Trans. Electron Devices* **36,** 1456 (1989).
27. S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the Universality of Inversion Layer Mobility in Si MOSFETs: Part I—Effects of Substrate Impurity Concentration," *IEEE Trans. Electron Devices* **41,** 2357 (1994).
28. R. Pierret, *Field Effect Devices*, Addison-Wesley Publishing Co., Reading, MA, 1990.
29. Y. Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill Book Co., Inc., New York, 1987.
30. D. A. Antoniadis, I. J. Djomehri, and A. Lochtefeld, "Electron Velocity in Sub-50-nm Channel MOSFETs," *Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, September 2001, pp. 156, 161.
31. Y. Kanada, "A Graphical Representation of the Piezoresistive Coefficients in Silicon," *IEEE Trans. Electron Devices* **ED-29,** 64 (1982).
32. A. Lochtefeld and D. Antoniadis, "Investigating the Relationship Between Electron Mobility and Velocity in Deeply Scaled NMOS via Mechanical Stress," *IEEE Electron Device Lett.* **22,** 591-593 (2001).
33. H. Wong, D. Frank, and P. Solomon, "Device Design Considerations for Double-Gate, Ground-Plane, and Single-Gated Ultra-Thin SOI MOSFET's at the 25 nm Channel Length Generation," *IEDM Tech. Digest*, p. 407 (1998).
34. R.-H. Yan, A. Ourmazd, and K. F. Lee, "Scaling the Si MOSFET: From Bulk to SOI to Bulk," *IEEE Trans. Electron Devices* **39,** 1704 (1992).
35. D. Esseni, M. Mastrapasqua, C. K. Celler, F. H. Baumann, C. Fiegna, L. Selmi, and E. Sangiorgi, "Low Field Mobility of Ultra-Thin SOI N- and P-MOSFETs: Measurements and Implications on the Performance of Ultra-Short MOSFETs," *IEDM Tech. Digest*, p. 671 (2000).
36. L. Chang, S. Tang, T. King, J. Bokor, and C. Hu, "Gate Length Scaling and Threshold Voltage Control of Double-Gate MOSFETs," *IEDM Tech. Digest*, p. 719 (2000).

Received June 5, 2001; accepted for publication October 24, 2001

Anthony Lochtefeld *Amberwave Systems Corporation, 13 Garabedian Drive, Salem, New Hampshire 03079 (alochtefeld@amberwave.com).* Dr. Lochtefeld received the B.S.E.E. degree from Ohio Northern University (1990) and the M.S.E.E. degree from Purdue University (1996) while investigating optical and electrical properties of nonstoichiometric gallium arsenide. He received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (2001), exploring carrier transport in deep-sub-100-nm MOSFETs as well as process integration issues for alternative MOSFET architectures. Dr. Lochtefeld is currently involved in technology development for CMOS in Si/SiGe heterosystems at AmberWave Systems Corporation in Salem, New Hampshire.

Ihsan J. Djomehri *Microsystems Technology Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139.* Mr. Djomehri received the B.S. degree in electrical engineering and computer science and the A.B. degree in physics from the University of California at Berkeley in 1997. At the Massachusetts Institute of Technology in 1998, he received the S.M. degree while developing a novel nanofabrication technology. He is currently pursuing his Ph.D. degree in electrical engineering and conducting research with Professor Antoniadis on inverse modeling of sub-100-nm MOSFETs, sponsored by the SRC. Mr. Djomehri's professional interests include device technology and computational physics.

Ganesh Samudra *Microsystems Technology Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139.* Dr. Samudra received his M.Sc. degree from the Indian Institute of Technology, Mumbai, and his M.S., M.S.E.E., and Ph.D. degrees from Purdue University. From 1986 to 1989, he worked for Texas Instruments, where he was elected Member, Group Technical Staff, for outstanding technical contributions in process and device simulation. Dr. Samudra is an Associate Professor on leave from the Department of Electrical and Computer Engineering at the National University of Singapore and, currently, a Visiting Professor at MIT. He has published about fifty articles in international journals and conferences. His research interests involve the development and application of technology CAD.

Dimitri A. Antoniadis *Microsystems Technology Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139 (antoniadis@mtl.mit.edu).* Dr. Antoniadis received his B.S. degree in physics from the National University of Athens in 1970, his M.S.E.E. in 1973, and his Ph.D. in electrical engineering in 1976 from Stanford University. From 1969 to 1976, he conducted research in the area of measurement and modeling of the earth's ionosphere and thermosphere. Starting with the development in 1976 of the now industry-standard SUPREM process simulator, his technical activity has been in the area of semiconductor devices and integrated circuit technology. Dr. Antoniadis has worked on the physics of diffusion in silicon, thin-film technology and devices, and quantum-effect semiconductor devices. His current research focuses on the physics and technology of extreme-submicron Si and Si/SiGe MOSFETs; silicon-on-insulator (SOI) devices and technology for sub-100-nm CMOS; advanced device and interconnect technology (e.g., 3D integration) for high-performance CMOS; and technology CAD and applications for advanced device design. He has authored or co-authored approximately 200 technical articles. He has received the Solid State Science and Technology Young Author Award of the Electrochemical Society in 1979, and the Paul Rappaport Award of the IEEE in 1998. From 1970 to 1971, Dr. Antoniadis was a Fellow of the National Research Institute, Athens. From 1976 to 1978, he was a Research Associate in the Department of Electrical Engineering at Stanford University. In 1978, he joined the faculty at MIT, where he is the Ray and Maria Stata chaired Professor of Electrical Engineering. From 1984 to 1990, he was the founding Director of the MIT Microsystems Technology Laboratories. Currently he is Director of the multi-university Focus Research Center for Materials Structures and Devices. Dr. Antoniadis has been a Fellow of the IEEE since 1986.