Silicon CMOS devices beyond scaling

W. Haensch
E. J. Nowak
R. H. Dennard
P. M. Solomon
A. Bryant
O. H. Dokumaci
A. Kumar
X. Wang
J. B. Johnson
M. V. Fischetti

To a large extent, scaling was not seriously challenged in the past. However, a closer look reveals that early signs of scaling limits were seen in high-performance devices in recent technology nodes. To obtain the projected performance gain of 30% per generation, device designers have been forced to relax the device subthreshold leakage continuously from one to several nA/μm for the 250-nm node to hundreds of nA/μm for the 65-nm node. Consequently, passive power density is now a significant portion of the power budget of a high-speed microprocessor. In this paper we discuss device and material options to improve device performance when conventional scaling is power-constrained. These options can be separated into three categories: improved short-channel behavior, improved current drive, and improved switching behavior. In the first category fall advanced dielectrics and multi-gate devices. The second category comprises mobility-enhancing measures through stress and substrate material alternatives. The third category focuses mainly on scaling of SOI body thickness to reduce capacitance. We do not provide details of the fabrication of these different device options or the manufacturing challenges that must be met. Rather, we discuss the fundamental scaling issues related to the various device options. We conclude with a brief discussion of the ultimate FET close to the fundamental silicon device limit.

1. Introduction

The tremendous success of CMOS technology is due to the scalability of the MOSFET transistor. Over two decades, very little has changed in the basic transistor design. A potential barrier controlled by the gate field modulates the current flow from source to drain. Its simplicity, together with the fact that it is available in complementary n-FET and p-FET versions, is the underlying basis for the success of CMOS technology. Questions about the end of scaling have been raised many times, but engineering ingenuity has repeatedly proven the predictions wrong. The most spectacular failures in predicting the end involved the "lithography barrier," in which it was assumed that spatial resolution smaller than the wavelength used for the lithographic process (~400 nm) is not possible [1, 2] and the "oxide scaling barrier," in which it was claimed that the gate oxide thickness cannot be reduced below ~3 nm because of catastrophic gate leakage [3, 4]. Furthermore, there was a substantial discussion on transport in MOSFETs when the deep-submicron regime gate length was reached, involving expectations that non-equilibrium effects, such as velocity overshoot, would enable greater gains in performance than expected from conventional scaling [5]. There is little evidence in the data to suggest that the MOSFET design of 2006 behaves in a fundamentally different manner than it did two decades ago. Scaling theory [6] gives us a recipe for increasing transistor performance; however, within the possibilities of technology, it is becoming increasingly difficult to meet transistor performance gains with reasonable device leakage.

Since we have been able to break through several "brick walls," now that we have devices in production that measure several tens of nanometers in gate-length dimension, the question can be reversed: Can we expect transistor performance to increase forever? The answer to this question challenges device designers and technologists, both of whom seek solutions that go beyond conventional scaling. We are now in an area

0018-8646/06/$5.00 © 2006 IBM

Table 1 Relationships for constant-field scaling and for generalized field scaling. α is the scaling factor for dimensions, and $E = V/\alpha$ is the normalized electric field.

Parameter	*Constant-field scaling*	*Generalized field scaling*
Physical dimensions, L, W, T_{ox}, wire pitch	$1/\alpha$	$1/\alpha$
Body doping concentration	α	E/α
Voltage	$1/\alpha$	E/α
Circuit density	$1/\alpha^2$	$1/\alpha^2$
Capacitance per circuit	$1/\alpha$	$1/\alpha$
Circuit speed	α	α (goal)
Circuit power	$1/\alpha^2$	E^2/α^2
Power density	1	E^2
Power–delay product (energy per operation)	$1/\alpha^2$	E^2/α^3

in which it is no longer sufficient to simply scale the dimensions of the device. Material properties set a natural boundary for what is possible. The permittivity constant of the gate insulator and the mobility of the channel material (wafer substrate) have not (or have only slightly) participated in scaling. In particular, the thickness of the SiO_2-based gate dielectric is a serious limiter of further scaling. Data shows that gate tunneling has become a major concern at about 1 nm gate dielectric thickness [7, 8]. Channel mobility in MOSFETs is trending toward lower values due to higher vertical fields [9, 10]. Engineering effort and physical understanding are directed to address the material questions. Gate dielectric research is seeking materials with a larger dielectric constant [8, 11, 12], and techniques to increase channel mobility through stress or substrate engineering are well underway [13–16]. With the right material solutions, MOSFETs will progress to the 10-nm-gate-length regime.

More recently, the end-of-scaling question has been raised from the perspective of energy dissipation on the chip. Energy dissipation was previously related only to active power. In current high-performance technologies, passive power contributes a significant part of the power balance. To contain passive power, voltage scaling has slowed, preventing widespread use of power-supply voltages at much below 1 V. Given that there is a fixed amount of power per chip that can be removed, tradeoffs must be made between active and passive power, which can have a detrimental effect on chip performance. Package and architecture solutions can help [17, 18], but ultimately the power vs. performance tradeoff will force us to look for a different way to design and use devices.

The organization of this paper is as follows. In Section 2 we review what scaling had to offer in the past and where it breaks down. In Section 3 we revisit scaling under energy constraint and discuss some device design tradeoffs. In Section 4 we discuss particular device design questions related to new gate materials, device structures, and substrate materials. Finally, in Section 5 we attempt to address how far silicon-based devices can be pushed, and the elements that restrain us from going further. Conclusions in Section 6 close the paper.

2. Scaling

This section briefly reviews some of the basic principles of scaling. It shows the benefits when scaling works well and also how such benefits are greatly diminished in the present era, in which the power-supply voltage V is not scaled below about 1 V for high-performance processors. In **Table 1** we present the scaling rules for both the original constant electric field and a generalized case. The former case, in which voltage is scaled down in direct proportion to physical dimensions, is described in [6]; the latter case, in which the electric field is allowed to be an independent variable E, defined as V divided by the dimensional scaling factor α, is presented in [19].

The simple concept of scaling for MOS transistors is to reduce all of the physical dimensions by the same amount α, while increasing the body doping and reducing the applied voltage to cause the depletion regions within the devices to scale as much as the other dimensions. Progress in microelectronics is also linked to scaling of the wiring dimensions, particularly the wiring pitch. For simplicity it is assumed in this discussion that wiring pitch is scaled by the same factor α used in the device, as has generally been the practice. (It has been shown that when devices and wires are scaled independently by different factors α_d and α_w, the speed is predominantly determined by α_d and the circuit density by α_w, whereas all of the important power parameters are affected by both [20].)

A first important result of scaling is the increased circuit density. This was seen in the early days as a key to reducing manufacturing costs, but over the years it has changed the whole shape and course of computing. A second important result that underlies the speed and power benefits is the reduction of capacitance per circuit, which can be understood as being due to the reduction of transistor widths and wire lengths, with the capacitance per unit dimension (e.g., C/μm) remaining essentially unchanged by ideal scaling. (This ignores the trends toward thicker wires and low-k insulators between the wires, which tend to offset each other.)

Another very significant benefit of scaling has been higher speed. In constant-field scaling, which is largely associated with the early n-MOS work, it is easily shown that circuit speed should increase directly with the

amount of scaling α. In CMOS technology over the last ten years, it has proved to be impossible to scale V and maintain speed increases because of constraints on the threshold voltage in order to avoid rising standby power in the "off" transistors. In this era (it is asserted here) the electric field E has been steadily increased by scaling V by less than α in order to meet the goal of increasing circuit speed by α, as outlined in the last column of Table 1.

It is well known that constant-field scaling provides much lower power per circuit, constant power density, and a power–delay product (energy per operation) which improves by α^3. As shown in Table 1, all of these are multiplied in generalized field scaling by E^2; it is no wonder that power and power density are now a major concern.

Unfortunately, we are now in an era in which voltage is not being scaled at all for a given application. Therefore, the parameter E rises directly with α, and we find circuit power constant with scaling, power density rising as α^2, and the power–delay product improving only by α. With respect to the power and power density, of course, it is assumed that the circuit speed actually increases with scaling, which is becoming very difficult to achieve.

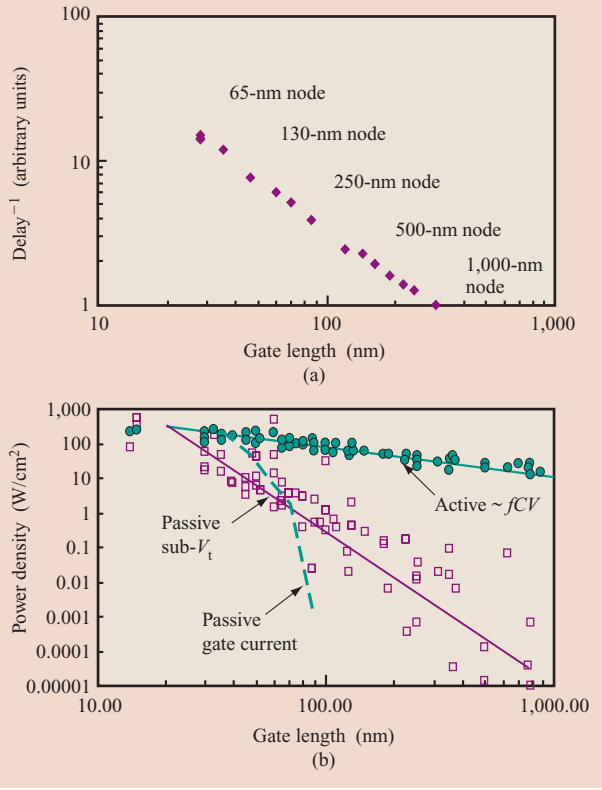
3. Power-constrained device scaling

While scaling has enabled decades (both in time and scale) of improvement in CMOS VLSI, the rapid growth in subthreshold leakage has finally, fundamentally altered the direction of power/performance improvements to CMOS technology [21, 22]. **Figure 1(a)** shows the improvements in intrinsic transistor delay to the power of −1; **Figure 1(b)** illustrates the growth of active and passive power density with scaling from 1-μm CMOS to 65-nm CMOS technologies. A significant transition occurs in the 130–65-nm regime, where passive power density moves from a minor part of the total to becoming dominant. These results have effectively halted traditional scaling in CMOS.

Leakage-limited drive current and gate length

It is of interest to explore the consequences with respect to gate length for a subthreshold-leakage-limited transistor design. Typically device comparisons have featured $I_{dsat} = I_{ds}(V_{gs} = V_{ds} = V_{dd})$ vs. I_{off} as a measure of leakage-limited transistor speed. Here I_{dsat} is the drain-to-source current I_{ds} at gate-to-source voltage V_{gs} and source-to-drain voltage V_{ds} at nominal power-supply voltage V_{dd}. The off-current I_{off} is measured at $V_{ds} = V_{dd}$ and $V_{gs} = 0$ V. It has been shown that a significantly more accurate predictor of CMOS inverter delay is given [23] by

$$I_{eff} = (I_{high} + I_{low})/2, \quad (1)$$



ERROR: See p. 361A.

Figure 1

(a) MOSFET performance vs. gate length; normalized MOSFET intrinsic device delay (CV/I_{eff}) vs. gate length. (b) Power density vs. gate length; data collected from literature for active power density and passive power density. Lines are intended to show trend. (fCV = frequency × capacitance × voltage.)

where $I_{high} = I_{ds}(V_{gs} = V_{dd}, V_{ds} = V_{dd}/2)$ and $I_{low} = I_{ds}(V_{gs} = V_{dd}/2, V_{ds} = V_{dd})$. The upshot of this exercise is that the effective drive current is sensitive to short-channel effects in a manner that $I_{dsat} = I_{on}$ is not. I_{high} is sampled at $V_{ds} = V_{dd}/2$ and is lowered from I_{dsat} by the drain output conductance, which is a short-channel effect that is strongly dependent on drain-induced barrier lowering (DIBL). The I_{low} term is more sensitive to V_{tsat} than I_{dsat}, since it is sampled at $V_{gs} = V_{dd}/2$; thus, the sensitivity to subthreshold swing degradation is amplified over that of I_{dsat}.

For a fixed I_{off}, as the gate length L_g is decreased, the longitudinal electric field driving the channel current is increased. However, I_{low} is decreased by two factors that increase V_{tsat}: increased subthreshold swing and the $1/L$ term in I_{off}. I_{high} is decreased by these same two factors, and additionally by increasing drain conductance. The net result is that for a given device structure, I_{eff} increases with decreasing L_g up to the point at which the discussed

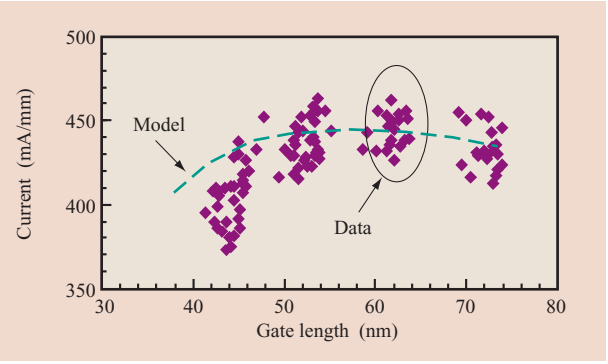


Figure 2

Effective switching current (I_{eff}) vs. gate length (L_g) model and data comparison. The maximum is a result of degraded short-channel behavior for shorter gate length.

short-channel effects finally dominate, and I_{eff} *decreases* thereafter. **Figure 2** shows an example of I_{eff} calculated for a fixed I_{off} (40 nA/μm) vs. L_g compared with experimental data.

Power vs. power density

Two views must be encompassed when examining power-related issues in CMOS scaling: one of total power, or power per circuit, and a second of power density, or power per unit area. It was seen earlier that classic scaling leaves power density fixed and, because of the quadratically decreasing area per circuit, quadratically decreases the power per circuit. Thus, two metrics, performance per power per circuit (F/P/Ckt) and performance per power density (F/Pd) are examined for scaling. In **Figures 3(a)** and **3(b)**, these two metrics are plotted vs. L_g, from the same data used to construct Figures 1(a) and 1(b). While the performance (F) trend has been continued beyond 130 nm ($L_g \sim 70$ nm), the growth rate in F/P/Ckt has dropped from a (classic scaling) $\sim L^3$ to below $\sim L^2$ as we enter the 65-nm node ($L_g \sim 35$ nm). Thus, the benefit to CMOS VLSI in terms of speed per power-function is dropping from a cubic dependence to a much slower improvement rate. More significantly, the improvement in F/Pd has not only slowed, but actually reversed to become degraded; that is, for functions that may be power-density-limited, design innovation is required in order to avoid an increase in power density, even with no increase in circuit speed!

Transport vs. electrostatics

A trend that may have begun in the 90-nm node shows only modest gains in short-channel scaling, aided by improvements in doping profile advances and limited to modest decreases in effective dielectric thickness T_{ox} [24]. Most of the performance gain for the 90-nm node and beyond comes from electron and hole mobility enhancements. In **Figure 4** we show three scaling cases. In the first, no new structural innovations are introduced, and the mobility of transistors is held fixed at the values achieved in 65-nm CMOS. This case shows no significant improvement in short-channel scaling. The second case assumes the same structural assumptions as in case 1, but the mobilities of electrons and holes are presumed to increase by 1.5x per generation. In the third case, in addition to the mobility assumptions of case 2, structural innovations are introduced at each node to improve control of short-channel effects, effectively keeping DIBL and swing constant at each node while gate length is reduced in concert with the node. (These structural innovations might be, for example, the introduction of double-gate transistors with reduced body thickness.) In all cases the power-supply voltage was kept fixed at 1 V, as was the effective gate dielectric thickness. In the first scenario, minor improvements in junction technology are overwhelmed by intrinsic short-channel effects associated

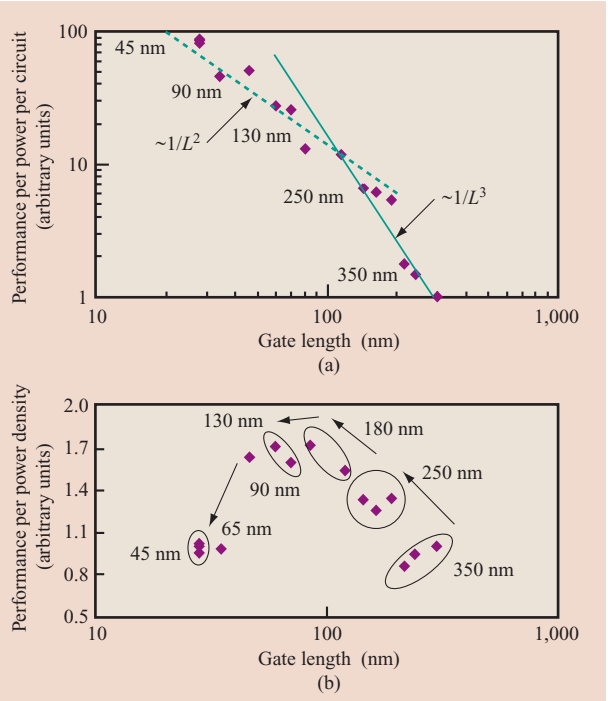


Figure 3

(a) Performance per power per circuit vs. gate length. Data from Figure 1(b) used to calculate performance (frequency) normalized to power and circuit count with respect to scaled density. (b) Performance per power density vs. gate length. Slowing of voltage scaling causes a reversal of the trend beyond 130-nm-node technology.

with shrinking gate lengths, and the degrading effects on effective drive, illustrated in Figure 3(a), outpace the performance gains from reduced gate capacitance; a net loss in transistor speed results.

In the second scenario, significant mobility increases, achieved for example through strain or improved channel materials, are able to boost current drive initially; however, the same short-channel factors in the effective drive eventually dominate and reverse the performance trend by the 25-nm node. When short-channel effects are also improved at a rate approaching the shrink factor, as in the third scenario, the transistor speed avoids degradation until it reaches the 25-nm node, although, even in this case, the performance gain is very marginal. From this model study one may conclude that significant innovations in both enhancements to transport and short-channel effect suppression are necessary for continued advancement of CMOS speed when power is constrained.

Finally, it is important to consider the choice of V_{dd} in the power-constrained era. In **Figure 5** transistor speed per power is plotted vs. transistor speed for a fixed $L_g = 35$ nm, 65-nm CMOS technology. Choices of V_{dd} from 0.8 V to 1.1 V are shown, and for each the off-current is varied from 1 nA/μm to 1 μA/μm. An envelope of best design points is sketched on the basis of these curves. It can be seen that there is a direct tradeoff of transistor speed for power efficiency.

As greater speed is demanded, one can either decrease V_t or increase V_{dd}. When V_t is decreased, the power eventually becomes dominated by subthreshold leakage; hence, the speed/power efficiency eventually degrades exponentially. At that point, increasing V_{dd} poses a more favorable return of speed per power. Hence, when comparing power vs. performance efficiencies of different transistor options, one must take care that comparable optimization has been established for all of the cases under consideration.

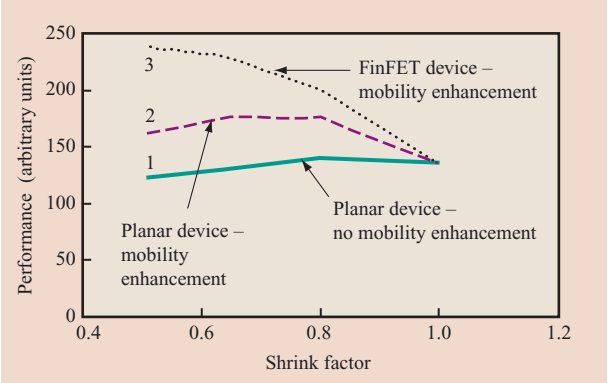


Figure 4

Performance vs. gate length shrink factor, for constant V_{dd}. Data is normalized to 65-nm-node planar device technology. Curve 1: only gate-length scaling; curve 2: gate-length scaling and mobility improvement; curve 3: gate-length scaling, mobility, and structure-enabled short-channel-effect improvement.

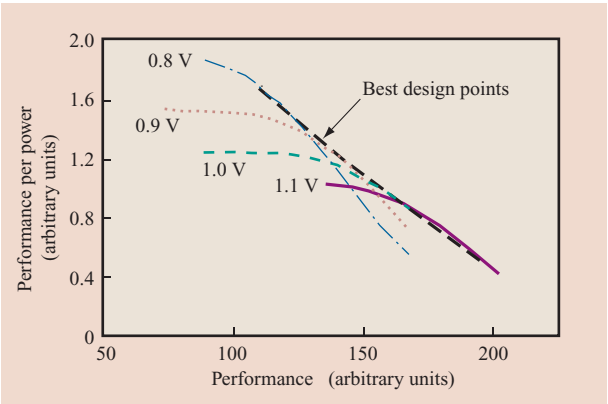


Figure 5

Performance vs. performance per power. For a given V_{dd}, I_{off} is varied (V_t setting) from 1 nA/μm to 1 μA/μm. Solid curve: 1.1 V, dashed curve: 1.0 V, dotted curve: 0.9 V, dashed-dotted curve: 0.8 V.

4. Device scaling

Short-channel effects

A successful device design delivers maximum on-current (I_{on}) at an acceptable device off-current (I_{off}), which has constantly increased in the previous technology generations to keep up with the device performance requirements. Given a constant supply voltage, high drive current would require a low threshold voltage, which is contrary to the desire for a low I_{off}. In addition to the threshold voltage (V_t), the subthreshold swing (S) also determines the device off-current:

$$I_{off} \propto 10^{-(V_t/S)}. \quad (2)$$

Threshold-voltage reduction with gate length and subthreshold swing are related to the device structure. In contrast to the on-current, they are only weakly dependent on the transport properties. They are related to the electrostatic behavior of the device. The subthreshold swing is determined by the gate modulation of the potential barrier height ($\theta_{barrier}$) that separates the source from the drain. For a partially depleted doping-controlled device, this would be the capacitance divider between gate dielectric C_{ins} and channel depletion width C_{depl}:

$$S = 2.3\frac{kT}{q}\left(\frac{\partial\theta_{barrierg}}{\partial V_g}\right)^{-1}. \quad (3)$$

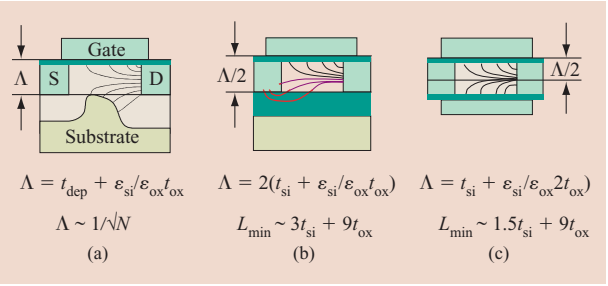


Figure 6

Scaling potential of different device types: (a) bulk; (b) FDSOI; (c) FD double-gate device. L_{min} estimates are done assuming a gate oxide dielectric and Si substrate and $L_{min} = 1.5\Lambda$. From [25], reproduced with permission; ©2001 IEEE.

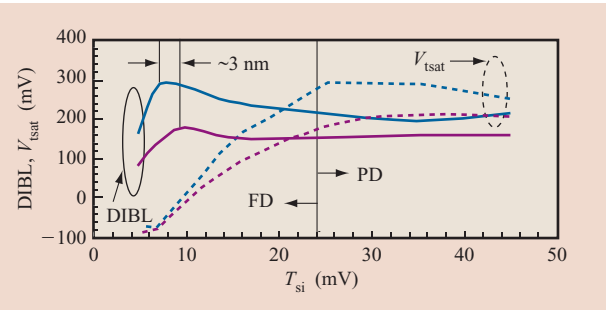


Figure 7

n-FET T_{si} scaling with fixed channel doping. Solid curves: V_{tsat} at L_{nom} and $V_{ds} = 1$ V; dashed curves: DIBL at L_{min}. Blue curves: short-device $L_{nom} = 28$ nm ($L_{min} = 25$ nm); red curves: long-device $L_{nom} = 46$ nm ($L_{min} = 38$ nm). (FD = fully depleted; PD = partially depleted.)

Consequently, a partially depleted doping-controlled device cannot achieve the ideal subthreshold swing of approximately 60 mV/dec. This capacitance divider is not present in a fully depleted double-gated device in which the front and back gates are connected to the same potential. Therefore, this type of device has the potential to achieve the best subthreshold swing.

A third short-channel effect is DIBL. This phenomenon is due to the modulation of $\theta_{barrier}$ with the drain voltage. It is a measure of the number of field lines originating from the drain that terminate at the source side of the channel. DIBL modulates the threshold voltage with respect to drain-to-source voltage and affects the effective drive current I_{eff} [Equation (1)]. In **Figure 6** we compile the scaling behavior for different device architectures as obtained from a generalized scaling theory [25]. The electrostatic scaling length Λ is a measure of how the scaling behavior of the device is related to various device properties, such as gate length L_g, channel doping N, body thickness T_{si}, gate dielectric thickness T_{ins}, and gate dielectric ε_{ins}. Shrinking gate length for partially depleted doping-controlled devices requires an ever-increasing doping level in the device, which may aggravate subthreshold, swing, DIBL, and junction leakage, as we later see. Ultimately, the threshold voltage of a fully depleted (FD) device is set by the body thickness of the device. Owing to a better shielding of the drain and source fields, the double-gate device, at a given minimum gate length, requires a less stringent body thickness than a fully depleted SOI device. In cases in which T_{si} is not significantly smaller than the minimum gate length, doping in the body is needed to control the short-channel behavior, although the body can remain fully depleted. The general scaling theory does not capture this situation accurately. In the following we investigate the transition from a partially depleted, doping-controlled device to a body-thickness-controlled device and its impact on short-channel effects.

Single-gate partially depleted, doping-controlled devices

Leaving general scaling theory, we turn to some interesting results that were obtained with 2D process and device simulation [26]. Our simulations also include mixed-mode simulations of delay chains to study ac (switching) behavior. If not stated otherwise, these are fan-out-of-1 delay chains, with or without appropriate wire loads. We first investigate the transition of a doping-controlled partially depleted (PD) SOI device to a geometry-controlled fully depleted (FD) SOI device. For all practical purposes, the partially depleted SOI device behaves with respect to short-channel scaling quite similarly to the bulk device, except that charge can accumulate in the body and modify its characteristics (floating-body and history effects). Device history is discussed in a different contribution in this issue [27]. To illustrate the impact of body thickness scaling, we show in **Figure 7** the saturation threshold voltage for a polySi-gate n-FET device vs. body thickness at 46-nm and 28-nm nominal gate lengths. Also shown is DIBL for the same devices at corresponding minimum channel lengths of 38 nm and 25 nm, respectively. To better understand the influence of body scaling, we adjusted the halo implants to meet the leakage targets for the devices with the thickest bodies and subsequently thinned the bodies. Equivalent oxide thickness is fixed at 1 nm (physical gate dielectric thickness). We find three distinct regions in this study: the partially depleted device for $T_{si} > 25$ nm, the fully depleted device for $T_{si} < 25$ nm, and the body-thickness-controlled device for $T_{si} < 7$ nm. Owing to the higher halo dose for the shorter device,

we see a higher V_{tsat} and DIBL for the PD region, which is largely independent of T_{si}. The FD device shows a decrease in V_{tsat} which is proportional to the loss of doping due to the thinner body ($\Delta V_{tsat} \sim \Delta T_{si} \cdot N_{channel}$) and an increase in S. This is more pronounced for the shorter device and is related to an increase in drain-to-source coupling for the FD body. For $T_{si} < 7$ nm, the device is controlled by the T_{si} and is largely independent of the doping. We see (shift of maximum in DIBL) that for the shorter device, body control comes at a somewhat thinner body thickness. From the general scaling theory (no doping in channel), we find that L_{min}/T_{si} of about 5 is required for a T_{si}-body-controlled SOI device [Figure 6(b)]. The presence of doping in the channel will reduce this ratio, as we see shortly. In **Figure 8(a)** we show subthreshold swing vs. DIBL for a properly designed FD device at $T_{si} = 15$ nm compared with its PD counterpart for nominal ($L_g = 45$ nm and 28 nm) and minimum ($L_g = 42$ nm and 25 nm) gate length. Again we see the increase in DIBL by migrating to shorter channel length. We also observe a somewhat higher DIBL in FD devices. As we discuss later, the DIBL increase has a direct consequence on the performance level.

DIBL for the PD device is, of course, also modulated by device floating-body effect, which usually adds a constant contribution in addition to the short-channel effect. To evaluate the performance impact of T_{si} body scaling, we have calculated ring oscillator delays to capture simultaneously competing ac and dc effects. From the previous analysis we have seen that a FD device shows slightly degraded short-channel effects if it is doping-controlled. However, the thinner body reduces the junction capacitance and therefore is beneficial. In **Figure 8(b)** we compare the ring oscillator delay for PD ($T_{si} = 48$ nm) and FD ($T_{si} = 15$ nm) vs. device leakage. The devices are leakage-matched at minimum gate length $L_g = 42$ nm and $L_g = 25$ nm, respectively. For both cases the nominal gate length is 3 nm longer than the minimum gate length (allowing for presumed manufacturing tolerances). The graph compares performance gain through gate length with performance gain through body thickness reduction. First, thinning the body results in 5% performance gain due to reduced junction capacitance. Second, reducing gate length from nominal 45 nm to 28 nm gives 14% performance gain for both FD and PD devices. The limited performance increase comes from the fact that the effective drive current is degraded by approximately 16% due to poorer short-channel effects for the shorter device, as indicated in Figure 2. Short-channel scaling must be improved to further improve the effective drive current, as discussed in the previous section. One possible solution is to back off from aggressive channel-length reduction and find an optimal short-channel/delay design point. Another solution for

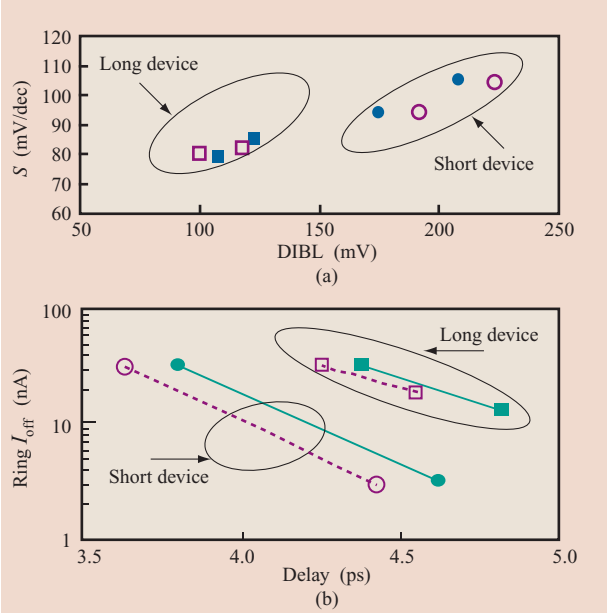


Figure 8

(a) DIBL vs. subthreshold swing for PDSOI and FDSOI devices; long devices $L_{nom} = 45$ nm ($L_{min} = 42$ nm); short devices $L_{nom} = 28$ nm ($L_{min} = 25$ nm). Solid symbols — PDSOI devices with $T_{si} = 45$ nm. Open symbols — FD devices with $T_{si} = 15$ nm. All devices have the same I_{off} at L_{min}. (b) Ring oscillator delay vs. ring oscillator leakage current. Devices are leakage-matched at $L_{min} = 42$ nm (boxes) and 25 nm (circles). Solid symbols — PDSOI devices with $T_{si} = 45$ nm. Open symbols — FDSOI with $T_{si} = 15$ nm.

increased drive current is to improve the gate coupling to the channel by using metal gates and high-k gate dielectric material. The benefit of the metal gate is the elimination of polysilicon depletion and thus increased gate control of the channel potential. The device off-current, together with the metal-gate workfunction, determines the short-channel behavior of the device at a given dielectric thickness. In **Figure 9** we show how the subthreshold swing and DIBL are modulated by the choice of metal-gate workfunction under constrained I_{off} at a minimum channel length of 17 nm. Halo doping is adjusted to meet the off-current at the minimum device length, and gate length is varied throughout the trajectory. For both the PDSOI device with $T_{si} = 48$ nm and the FDSOI device with $T_{si} = 10$ nm, a significant modulation of short-channel effects is observed with varying workfunction. Only for the FDSOI device with extremely thin body, $T_{si} = 5$ nm, the dependence of the short-channel effect on workfunction is weak. This is a direct consequence of the confinement of minority carriers in the channel by the physical thickness of the body, rather than the fields. The lower doping levels required for more mid-gap workfunctions reduce the vertical field and spread the

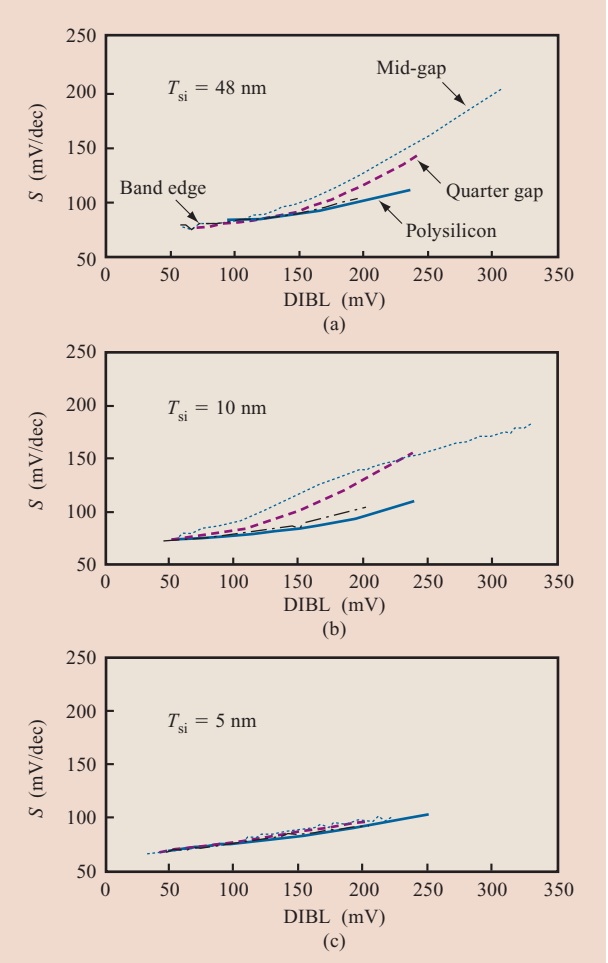


Figure 9

Subthreshold swing vs. DIBL for metal gate with different workfunctions and SOI devices with different body thicknesses. Gate length is varied: 17 nm, 22 nm, 30 nm, and 50 nm. I_{off} fixed at L_{min} = 17 nm by halo adjustments.

carriers into the body region during subthreshold operation. There is no shielding from drain to source, and the subthreshold swing is degraded. In **Figure 10** we show the carrier distribution at zero gate voltage for a device with a polysilicon gate and a device with a metal gate with a workfunction of 125 mV away from the band edge. Both devices are designed to have the same I_{off} and channel length. The study in Figure 9 suggests that an aspect ratio $L_{min}/T_{si} > 3$–4 can provide short-channel control that is independent of gate workfunction for a doping-controlled FD device.

To study the performance advantage of a metal-gate device, we calculated the delay of an inverter chain. In addition to varying the workfunction we also included in this calculation the sensitivity to high-k gate dielectric. To mimic the effect of a high-k gate dielectric, we increased the dielectric constant of the gate dielectric by a factor of 2, from 3.9 to 7.8, and kept the physical insulator thickness constant at 1 nm. We did not account for any mobility degradation [11, 12], in an effort to explore the best leverage attainable from these elements. In **Figure 11(a)** we show the delay for an unloaded ring oscillator of fan-out 1 for two different device lengths (L_{min} = 35 nm and L_{min} = 25 nm) and three metal workfunctions (quarter gap, band edge, and 110 mV away from the band edge) over the range of gate dielectric scaling (accommodated by changing the electrical permittivity at fixed dielectric thickness). For selected cases we also show the effect of an additional wire load to

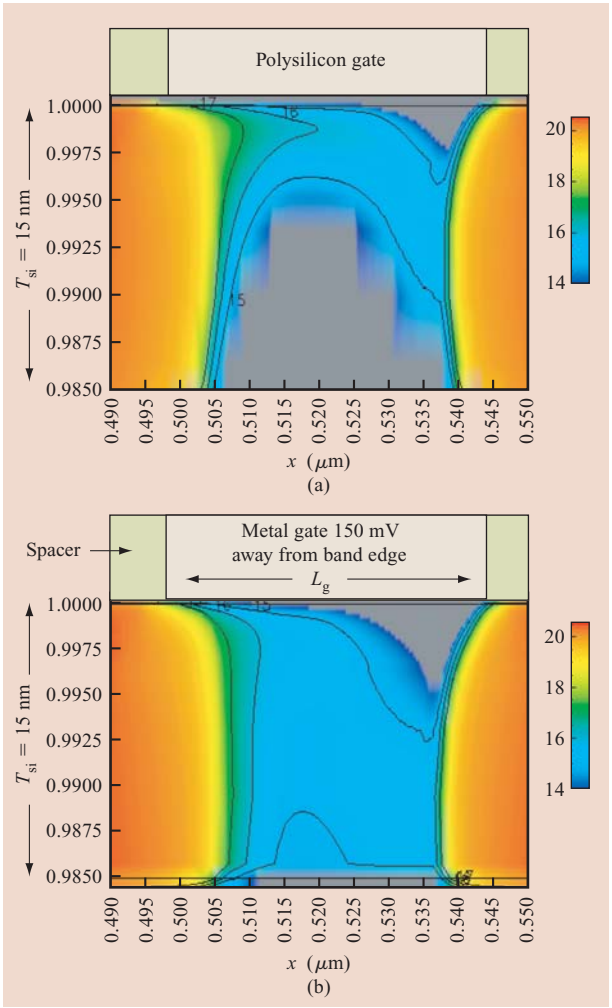


Figure 10

Carrier distribution in polysilicon-gate and metal-gate devices (150 mV away from band edge). Both devices have the same off-leakage at 42-nm gate length. The device shown has 45-nm gate length.

mitigate the effect of increased gate capacitance on ac performance. The delay is normalized to a polysilicon-gate device with the same gate length, and off-leakage current. The polysilicon-gate device has T_{inv} of 17.5 A for a 1-nm equivalent oxide thickness. Figure 11(a) clearly shows that for a quarter-gap workfunction, only for an extremely scaled dielectric does the metal-gate performance improve with respect to polysilicon-gate devices. The band-edge and close-to-band-edge workfunctions show almost equivalent behavior. The results indicate that in the gate length, off-leakage, and dielectric scaling space, the band-edge workfunction is not necessarily optimal. This behavior is due to the off-current constraint, since the band-edge metal requires higher doping in the channel, and this results in a mobility degradation which in turn degrades the drive current. The figure also shows that for currently achievable high-k dielectrics ($T_{inv} \sim 14$–15 A), ac performance gain for gate-loaded circuits is of the order of 5% to 10%. The gain is higher for partially wire-loaded circuits. As discussed earlier, the position of the workfunction has a significant impact on the short-channel behavior of the device because of the off-current constraint. In **Figure 11(b)** we compare the workfunction scaling behavior of high-performance (PDSOI) and low-power devices (bulk). The off-current leakage specification of the low-power devices is typically three orders of magnitude lower than that of high-performance devices. Therefore, the channel doping concentration (halo) is significant higher for the low-power device. To adjust for an off-band-edge workfunction, the halo dose must be reduced in order to maintain the off-current leakage for maximum drive. However, there is still enough doping left to guarantee carrier confinement, so that the impact on short-channel behavior is less sensitive to workfunction than in the high-performance case. Thus, gate metals with workfunctions around ±250 mV off mid-gap are useful and even beneficial for low-power devices, primarily because of improved mobility. Improved dielectric scaling gives a substantial performance gain for low-power devices. Assuming current high-k gate stack materials, a significant reduction in gate length results in substantial performance gains, as shown in Figure 11(b).

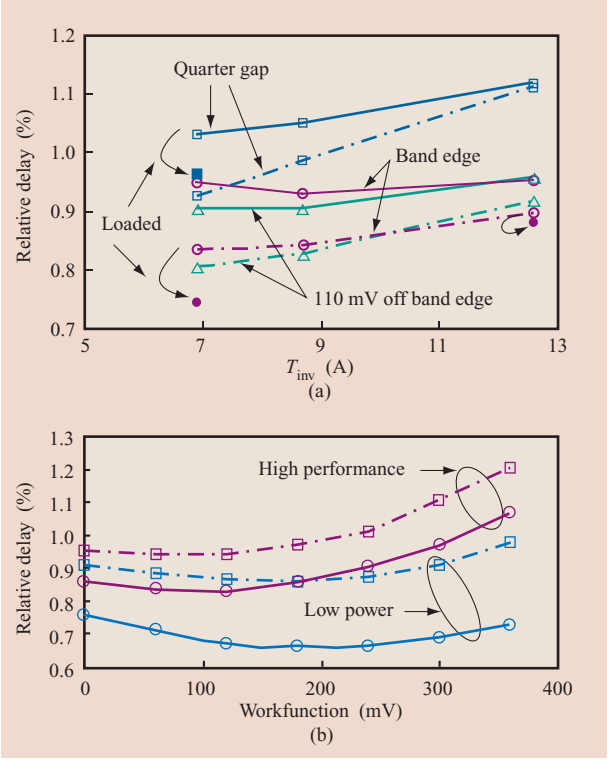


Figure 11

(a) Ring oscillator delay vs. dielectric scaling and metal gate workfunction, for L_{min} = 35 nm (solid curve) and L_{min} = 25 nm (dashed-dotted curve) for PDSOI devices with T_{si} = 48 nm. Workfunctions: open squares — quarter gap ±250 mV off mid-gap; open triangles — ±110 mV off-band edge; open circles — band-edge metal. Solid symbols account for the effect of wire load for the short device. (b) Loaded ring delay vs. workfunction for high-performance and low-power devices. Band-edge workfunction corresponds to 0 V. Open boxes — high-performance devices: L_g = 37 nm, T_{inv} = 21 A; open circles — low-power devices: L_g = 42 nm, T_{inv} = 28 A (low gate leakage!); dashed-dotted curves — polysilicon gate replaced by metal gate; solid curves — high-k dielectric.

Multi-gate devices

Multi-gate devices exhibit a scaling advantage due to better gate control of the channel [28–30]. Several versions of multi-gate devices are discussed extensively in the literature. There are two varieties: planar and vertical structures. The former group contains ground plane and back-gate devices, which are derivatives of the SOI device. The vertical structures contain the FinFET [31–34], Omega FET [35], and the Tri-Gate [36]. A dual-gate device can be used in two modes. In the first mode, the two gates are connected and move simultaneously. This configuration provides the best short-channel scaling advantage because of tighter gate control of the channel and superior subthreshold swing. A generalization of this type of operation to multiple-gate structures such as the Tri-Gate or Omega FET is straightforward. Operating a dual-gate device with independent gates introduces an additional gate-to-gate capacitance, but it also gives the opportunity to control the front-gate threshold voltage with the back gate. In this sense, it is comparable to a well-biased bulk device, with the advantage, however, that a much larger threshold-voltage swing can be accomplished with proper device design. We first turn to a comparison of multiple-gate and FDSOI devices. For

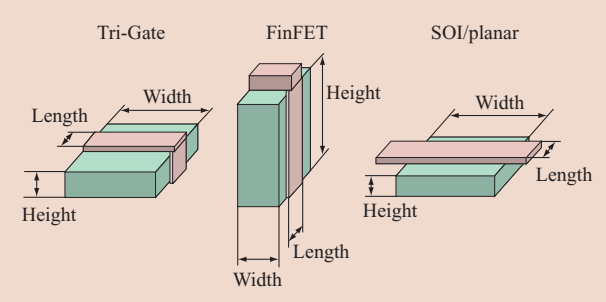


Figure 12

Geometry definition of multi-gate devices: Tri-Gate — three conducting surfaces; FinFET — two conducting surfaces (top surface not active).

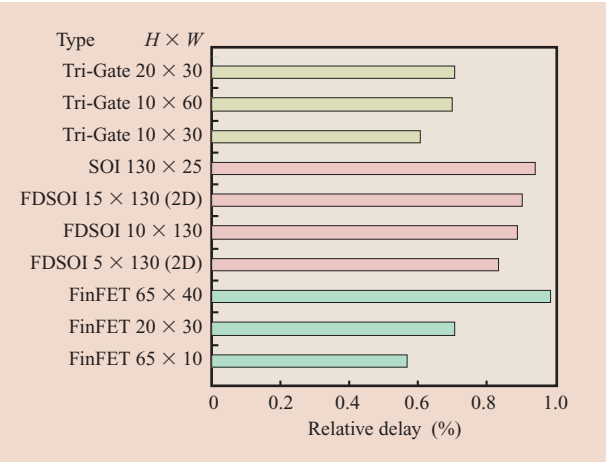


Figure 13

Relative delay for an unloaded fan-out-of-1 ring oscillator, normalized to a PDSOI device of same length. Devices have the same I_{off}, equivalent oxide thickness of 1 nm, and are operated at 1 V. $L_g = 23$ nm.

clarification, in **Figure 12** we define the geometry of the different device types.

The Tri-Gate is essentially a mesa-isolated SOI device in which the gate wraps around the active silicon area. All surfaces, sides of height H, and top surface of width W, contribute to channel conduction. The conduction in the corner, where vertical and horizontal surfaces meet, is an essential part of this device. The FinFET has a higher aspect ratio than the Tri-Gate device; the top surface is usually covered by a thick oxide and does not contribute to the channel conduction. Its height and width define the Fin. The height of the planar device is defined by the thickness of the active silicon. We have studied performance tradeoffs among these device types by keeping all parameters such as gate oxide thickness, doping profile, and series resistance constant. Of course, each type of device has its own optimization space. For the sake of comparison, we kept as much commonality as possible. All devices were constrained to the same I_{off} leakage at a given channel length, and their short-channel effect was doping-controlled. In **Figure 13** we show the relative change in ring delay compared with that of a PDSOI device with the same constraints. In this figure we have also included properly normalized 2D calculations. We see small performance advantages of FDSOI devices with respect to silicon body thickness, as discussed in detail above [Figure 8(b)]. More interesting are the Tri-Gate devices and the FinFET results. The improved performance of FinFET for thinner width (body thickness) is directly related to the better short-channel scaling behavior of the device, which allows lower threshold voltage (see Figure 4). Many of these advantages carry over to the Tri-Gate structure, even for a relatively open aspect ratio, as may be seen by comparing the 10×30 ($H \times W$) Tri-Gate, the 65×10 FinFET, and the 10×130 SOI structures in Figure 13. Much of this advantage is derived from the ability of the overlapping gates to partially screen the bottom of the SOI island from the drain field, which can easily penetrate through the thick buried oxide to affect the short-channel behavior of the planar SOI device. There is a slight advantage from increased mobility due to low surface field if the Fin is fully depleted. A potential advantage of the FinFET is that it can be operated with two independent gates, which offers a number of interesting benefits.

Although the FinFET offers improved scaling behavior, it may be a disruptive element with respect to circuit design, since the effective device width, and therefore the current, comes only in multiples of Fins. A planar double-gate device would have a continuous width but would suffer other drawbacks. One advantage for the FinFET is, for example, that both gates are self-aligned with the junctions so that the impact of overlap capacitance can be engineered by a proper junction design. For a planar back gate, the situation is not that simple, unless a back gate can be made to align itself with the front gate. The simplest back gate would be built on a SOI substrate with a thin buried oxide and a buried back gate. To operate the back gate in a reasonable voltage range, the body must be fully depleted and the buried oxide must be reasonably thin. In **Figure 14** we show the capacitance components for an unpatterned or non-self-aligned back-gated device. Although schemes exist to build self-aligned back-gated devices (with minimal additional capacitance) [37], they are usually difficult and expensive to implement. For an unpatterned back gate,

additional capacitance would be added through junction-to-back-gate coupling that would scale with the junction length. This capacitance could be mitigated by using a (non-self)-aligned back gate which, however, would also increase process complexity. For the choice of an unpatterned back gate, the proper performance design space is now an optimization of buried-oxide (back-gate dielectric) thickness, body thickness, and gate length, and the front-gate dielectric. **Figure 15** shows a result of such an optimization. In this figure we have normalized the delay impact to the SOI device (100 nm buried oxide) with the same body thickness. Again we have constrained the device off-leakage for all devices to the same value. The front-gate dielectric was 1 nm equivalent oxide thickness. For the back-gated devices we chose a grounded back gate for the n-FET and the back gate at V_{dd} for the p-FET. Halo doping was adjusted to meet the device off-current for the minimum device. The performance impact due to the back-gate dielectric scales linearly with its thickness, reaching 40% degradation for a 5-nm back-gate dielectric. We also see that the impact of the back-gate thickness does not depend significantly on the body thickness of the device, which allows the design point to be determined by other constraints. A reduction of junction-to-back-gate capacitance, as obtained by a patterned aligned back gate, reduces its capacitance penalty linearly with the back-gate-to-junction overlap. Reduction of the undesired back-gate capacitance for the unpatterned case can also be accomplished by reducing the doping level in the back gate. For the device with a 10-nm and 20-nm body thickness and 10-nm back gate, for example, we reduced the gate doping level from 10^{20} cm^{-3} to 10^{17} cm^{-3} and further to 2×10^{16} cm^{-3}. This of course also affects the effectiveness with which the back gate can be operated. However, this scheme represents an excellent tradeoff of back-gate control with process simplification afforded by the use of an unpatterned back gate. Regions where back-gate action is required would receive a highly doped back gate, and those where the performance impact is detrimental would be doped at a much lower level. In that scenario, performance devices would still be designed with halo doping and proper junction engineering. It would, however, provide an option to create SRAM devices without channel doping and set their thresholds by back-gate bias. This would eliminate one component in the threshold variation [38, 39] and would, in addition, allow device V_t values to be set appropriately at the end of the manufacturing process with appropriate control circuits. In **Figure 16** we show how two completely different devices can coexist on one wafer with essentially the same device structure. The logic device was optimized to have an I_{off} ratio of 10 between the nominal and the 6σ short device at a fixed

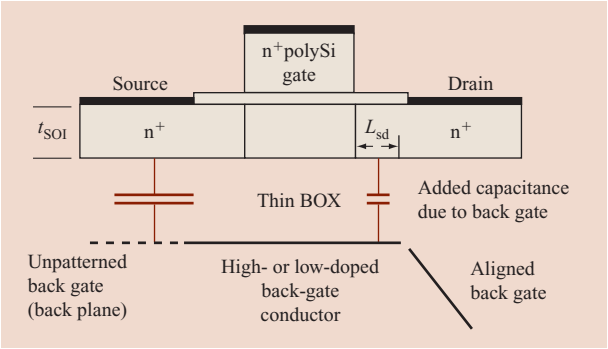


Figure 14

Capacitance components for an unpatterned (left) and aligned (right) back gate with diffusion-to-back-gate overlap L_{sd}.

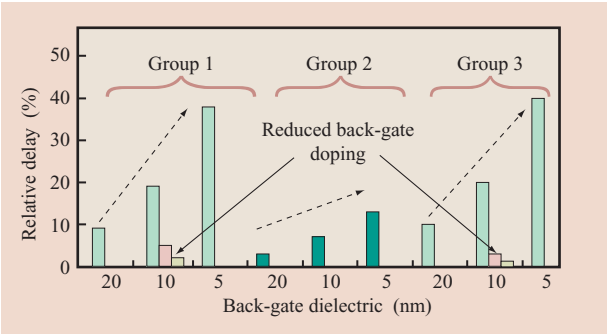


Figure 15

Performance impact of unpatterned and non-self-aligned back gate normalized to an FDSOI device of the same body thickness, channel length $L_g = 25$ nm, and I_{off}. Group 1: $T_{si} = 10$ nm, contact length 100 nm; Group 2: $T_{si} = 10$ nm, contact length 22 nm (aligned back gate); Group 3: $T_{si} = 20$ nm, contact length 100 nm. Dashed arrows – performance degradation with thinner back-gate dielectric; solid arrows – reduced junction-to-back-gate capacitance due to more lightly doped back gate.

leakage for the 3σ short device. Its short-channel behavior was adjusted with a proper halo implant and choice of junction design. The SRAM device did not receive the halo implant or a separate channel doping. With more negative back bias, we can bring the SRAM device with the undoped channel into an appropriate roll-off behavior without significant gate-length penalty.

We have discussed solutions and limitations of electrostatic scaling behavior to improve device performance. It was essential for this investigation that we went beyond the conventional I_{off}/I_{on} metric to measure the "goodness" of the device. We systematically applied ring oscillator delay calculation to compare the

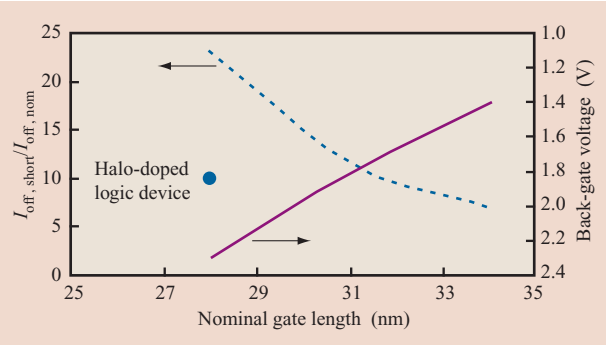


Figure 16

Leakage-limited design point for undoped body back-gate-controlled device, with 10-nm body thickness and 10-nm back dielectric thickness. Front gate is n^+ polySi and back gate is p^+ polySi. The nominal gate length is varied, and the ratio I_{off} short (L_{nom} − 6 nm) to I_{off} at L_{nom} is plotted (left side), with back-gate voltage plotted (right side) for constant leakage at a shorter (L_{nom} − 3 nm) device.

different device design options. We have learned that the best way to success is an improved gate dielectric scaling, as enabled with high-*k* dielectrics. For metal gates we do not see a benefit in performance if we cannot get the workfunction close enough to the band edge. Proximity of less than 100 mV to the band edge is required. We see a performance benefit for off-band-edge metals in the low-power application space. Device leakage still requires a substantial amount of doping in the channel, which provides for confinement of carriers to ensure good short-channel behavior. A device for which short-channel effects are independent of the metal gate workfunction requires a minimum gate length that is three to four times the body thickness at 1 nm equivalent oxide thickness. We also have shown that multiple-gate structures display some advantage over single-gate structures with respect to short-channel behavior. For vertical structures there is a better performance/complexity tradeoff through the dominance of corner current and sidewall shielding for Tri-Gate devices. For independent gates in multiple-gate structures, the FinFET has an advantage due to the self-aligned gates, which eliminate the additional capacitance that occurs in a planar structure with unpatterned or non-self-aligned gates. We have offered a solution to show how, with little process complexity, an unpatterned planar back-gate device can be used as a high-performance logic and SRAM device with essentially the same device structure.

High-mobility channels

Mobility is considered to be the key quantity in describing transport for MOS devices. We defer the discussion of mobility in the ultimate FET to later in the paper (Section 5) and assume that the devices under consideration are scattering-limited; it therefore makes sense to discuss the role of mobility as a means of enhancing device performance. The channel mobility in a FET has three distinctive regions [**Figure 17(a)**]. For low vertical fields or weak inversion, mobility is limited by Coulomb scattering due to doping atoms or charges at the gate dielectric/silicon channel interface. Moving to higher fields, phonon scattering dominates, and in still higher fields, surface roughness scattering becomes the limiting scattering mechanism for the channel mobility.

In circuit operation the device switching trajectory passes through various regions of the V_{ds}–V_{gs} space, and it is of interest to understand how these different scattering components influence performance. In **Figure 17(b)** we show the response of an unloaded ring oscillator to mobility change. We have examined the effect on performance of each scattering mechanism (Coulombic, phonon, interface roughness) separately by increasing only the corresponding mobility component [26, 40] by a factor of up to 2. Although we show results for only an inverter chain, similar dependencies were obtained for other circuit elements. We find that the phonon and Coulomb parts are similar in impact, whereas surface scattering has less effect. This is because the device spends less time in the high-gate field region (high V_{gs} and low V_{ds}) than in the other regions of the V_{gs}–V_{ds} space. In **Figure 17(c)** we show the performance impact over a wider range of total mobility variation for two different nominal device lengths constrained to the same off-leakage. We find that the relative performance impact depends only weakly on the gate length, with a strong tendency to saturate at higher mobility enhancements. The curve is calibrated to typical data for high-performance 65-nm-node devices. To obtain a comparable performance boost in future generations, mobility enhancements must be significantly larger than those achieved today. To have a significant impact on performance, mobility related to Coulomb and phonon scattering should be the focus of improvement. In a very simplistic picture, the Coulomb component could be improved by reducing the doping in the channel. Of course, this comes at the cost of short-channel degradation unless one takes advantage of a multiple-gate structure or a thin-body FDSOI device. From the analysis in the previous paragraph, we find that this would require, at a 28-nm gate length and an equivalent oxide thickness of 1 nm, a silicon body thickness of approximately 5–6 nm for an undoped channel, and 7–8 nm for a doped channel. Mobility degradation due to quantum confinement becomes significant [41, 42] at body thicknesses below 5 nm. Thinning the body further would greatly enhance the role of surface scattering and would

therefore further degrade mobility [43]. Multiple-gate structures are beneficial because the required body thickness for short-channel control is approximately twice that of single-gated FDSOI devices. For these body-thickness-related mobilities, degradation is not as significant, and sufficient short-channel behavior with reduced doping can be achieved. Two levers for the phonon-limited mobility are crystal orientation and a deformation of band structure by strain. For electrons,

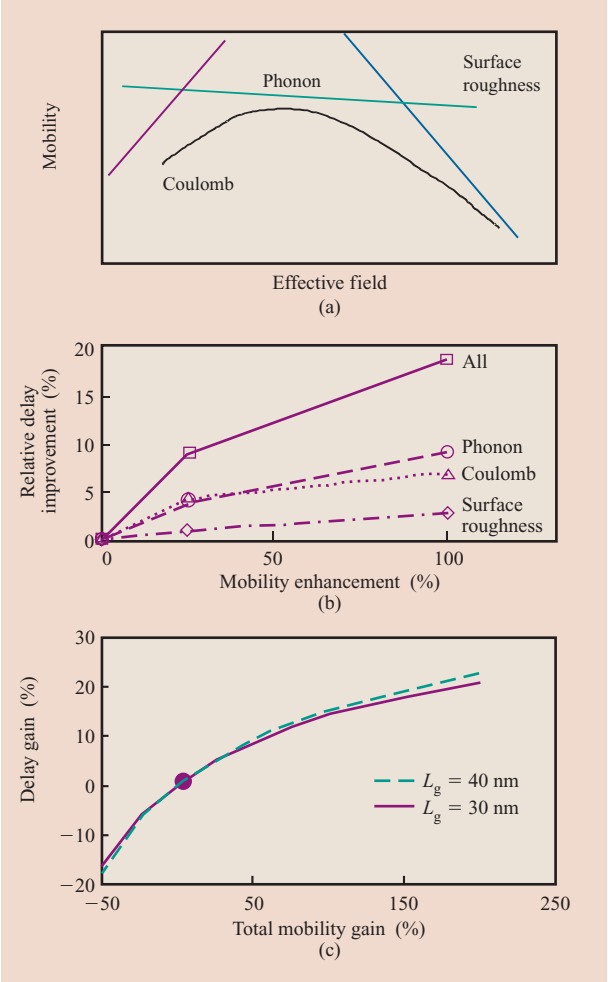


Figure 17

(a) Channel mobility components in a FET. (b) Impact of mobility components on ring oscillator delay. Dashed curve – phonon scattering; dotted curve – Coulomb scattering; dashed-dotted curve – surface roughness scattering. Calculation [26] was done for a 38-nm PDSOI device with T_{si} = 48 nm and equivalent oxide thickness of 1 nm. Calculation is based on calibrated Mujtaba mobility model [26, 40]. (c) Relative delay impact for an inverter delay chain with fan-out-of-3 vs. mobility enhancement. Model is calibrated to early 65-nm node (red dot). Devices with L_g = 30 nm and L_g = 40 nm have same I_{off}. No further device improvement assumed. Same mobility gain for n-FET and p-FET.

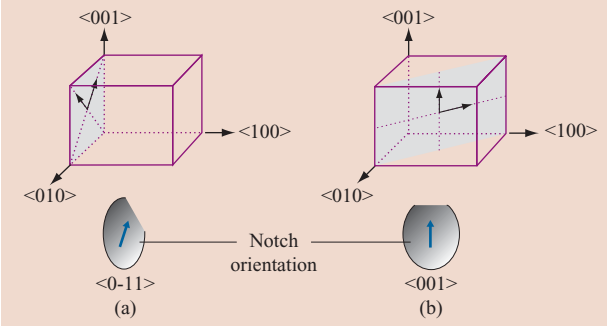


Figure 18

Crystal orientation (wafer surface: gray) and possible current flow strain directions: (a) (100) surface; (b) (110) surface.

strain in silicon splits the degenerate Δ states and lowers the energy of a sub-band with lower effective transport mass [13, 44] and high-density off-state mass, thus improving drive current. For holes, a complicated deformation of the bands can occur [45]. In addition, at higher strain levels scattering rates are also affected. Stress can be applied perpendicular to the wafer and in the plane of current flow. In that plane (wafer surface), it is convenient to separate stress in the direction of current flow and the direction perpendicular to that. Furthermore, the stress can be uniaxial, biaxial, compressive, or tensile. This leaves many different combinations, several of which have been realized [13–15, 46]. In **Figure 18** we show how wafer surface and in-plane direction line up. In **Figure 19** we show how the phonon-limited mobility behaves in this configuration space. Calculations were done using self-consistent linear response theory.[1] Scattering rates were calculated taking account of the silicon band structure for electrons and holes and phonon dispersion relations at a strain level of 1% compressive and tensile, respectively. In Figure 19(a) we show the results for electrons and in Figure 19(b) the results for holes. The change of mobility is normalized to the standard (100) surface in <110> directions on relaxed silicon substrates. First we see that for electrons the optimum surface is the standard (100) surface. For uniaxial strain, only tensile strain can provide some moderate gain in electron mobility. There is apparently no benefit for uniaxial compressive strain in the case of electrons. In general, these gains are independent of wafer and current flow orientations. Only for biaxial compressive strain can the electron mobility be improved in a more significant way if the wafer surface is (110) and the current direction is in the <0-11> direction. The situation looks more promising for holes. We see

[1]M. Fischetti, to be published.

Electron phonon-limited mobility

Uniaxial						
	Surface	Stress direction	Current direction	Rel.	Com.	Ten.
	100	100	100			
	100	100	010			
	100	010	110			
	011	100	100			
	011	100	0-11			
	011	110	100			
	011	110	110			
	011	110	1-10			
	011	110	11-2			
Biaxial						
	001		100			
	011		100			
	111		100			
	001		110			
	011		0-11			
	111		110			

(a)

Hole phonon-limited mobility

Uniaxial						
	Surface	Stress direction	Current direction	Rel.	Com.	Ten.
	100	100	100			
	100	100	010			
	001	110	110			
	001	110	-110			
	011	100	100			
	011	100	0-11			
	011	110	100			
	011	110	110			
	111	110	11-2			
	011	110	1-10			
Biaxial						
	001		100			
	001		110			
	011		100			
	011		0-11			
	111		11-2			
	111		1-10			

(b)

Severe degradation
Degradation
Little impact
Enhancement
Strong enhancement
Extreme enhancement

Figure 19

Phonon-limited mobility for (a) electrons and (b) holes for various crystallographic surfaces, current, and strain directions. Impact on mobility is normalized to standard surface (100) and <110> current direction and relaxed (rel.) substrate. Calculations are done for 1% tensile (ten.) and compressive (com.) strain. Mobility change quoted at inversion density 3×10^{12} cm^{-2} and 4×10^{12} cm^{-2} for uniaxial and biaxial strain, respectively.

significant mobility improvement, even in the relaxed case, if we select a different wafer surface than (100). A (110) wafer surface orientation with current flow in the <110> direction gives mobility that is more than two times higher than that in a standard wafer [16]. Further improvement can be obtained if the good surface orientation is subjected to compressive strain in <110> current flow. A combination of these two brings the hole mobility close to the electron mobility on standard wafers. Biaxial compressive strain for holes is best for the (110) surface. With sufficiently high levels of tensile strain, hole mobility also increases significantly on a (100) surface if current flow is restricted to the <110> direction [47].

Mobility is a long-channel property, and its impact is only indirectly measurable for short devices. The question at hand is this: If strain-enhanced mobilities can be implemented, how will they scale if the devices are scaled? In Figures 17(b) and 17(c), we show how ring-oscillator performance is affected by mobility improvement. The model shows a clear saturation of performance gain with higher mobility, which is the result of an early onset of velocity saturation. To achieve the same impact on performance from one generation to the next, the mobility gain must increase superlinearly. The hope is that this can be done with the combination of various stress techniques. The question remains, however, how much mobility gain is ultimately achievable by stress techniques? And how do these techniques scale with the technology [48–50]? In principle, we can distinguish between global-substrate-engineered stress and mechanisms that create a local stress field in the device. Strained-silicon technology [13, 47], for instance, is an example of the former. By growing a thin silicon device channel layer on a strained buffer material of a Si_{1-x} Ge_x composition, a channel mobility enhancement can be

engineered. Various techniques have been employed to generate local strain. It can be shown [51] that mobility enhancement due to biaxial substrate strain is only weakly dependent on channel length. Therefore, in the case of substrate strain, the mobility properties measured on a long-channel device correspond directly to the behavior of a short-channel device. However, it is important to notice that the relationship of long-channel mobility and short-channel behavior can be established only if self-heating in the short-channel device is taken into account because of the decreased thermal conductivity in Si_{1-x} Ge_x substrates.

In the case in which the stress is created locally, there is no relationship between the long-channel mobility and short-channel behavior of the device, since the strain field is usually self-aligned with the gate edges. The mobility for this second choice of local strain has to be inferred indirectly from the electrical data of the short-channel device compared with proper controls. The situation becomes even more complicated if there is a superposition of different strain sources [48–50]. For instance, these sources can be liner stress, which is mostly longitudinal with respect to the channel current flow; trench-isolation-induced stress, which can be both longitudinal and transverse to the direction of current flow (depending on the device width, for instance); or embedded SiGe, which is longitudinal to current flow. All of these components in general depend on details in the device layout and processing conditions, and can have either a cumulative or a compensating effect on the total relevant strain seen by the device. Therefore, engineering mobility enhancement by strain engineering can be a formidable undertaking and much more intricate than mobility enhancement by substrate engineering.

5. The ultimate (silicon) FET

Fundamental scaling limits for silicon devices

The question of the ultimate limits to FET scaling is an old one, dating back to the early days of MOSFET technology [1, 6]; it has been considered by many individuals over the years. In general, the predictions have been conservative, sometimes extremely inaccurate [1], stemming in general from a lack of appreciation for the range of applicability of certain limits, the scalability of the silicon dioxide insulator, and new structural possibilities such as ultrathin silicon-on-insulator or multiple-gate devices. Design enhancements, such as the use of laterally nonuniform channel doping (halo) have also contributed to the continuous progress of scaling.

An interesting subplot to the discussion of limits has been the question of electrostatic integrity: i.e., gates vs. drain voltage control of the internal potentials. A device with poor electrostatic integrity demonstrates, among other things, excessive DIBL. Dennard's scaling rules [6, 19] kept a constant electrical aspect ratio, and others [52] allowed a tradeoff between different parameters such as oxide thickness and junction depth, to maintain electrostatic integrity. In the 1990s, the subject of electrostatic scale length, or attenuation length, for drain-induced potential along the channel was discussed [25, 53] for both double-gated and bulk FETs. Some confusion has been caused by notation, since the scale length, Λ, is π times the attenuation length, so that Frank's criterion [25] of channel length greater than 1.5 Λ means ~4.7 times the attenuation length. For a double-gate FET, in the limit of zero thickness insulator, the scale length equals the silicon thickness (see Figure 6).

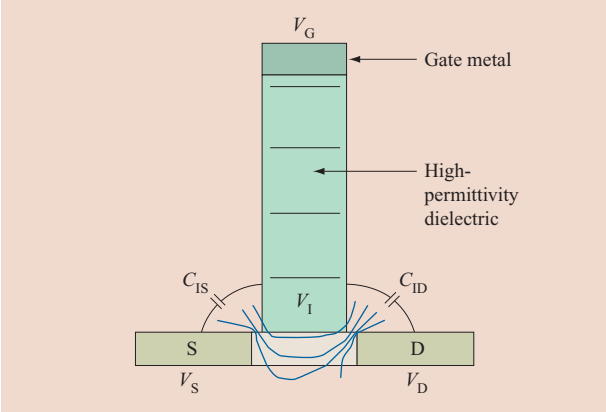


Figure 20

Ultrahigh-permittivity, high-aspect-ratio gate insulator showing sketched equipotentials (thin lines), and derivation of the potential, V_I, of the dielectric adjacent to the channel region.

High-permittivity insulators do offer a means of decreasing the electrical insulator thickness without decreasing the physical thickness for improved scalability. However, as pointed out by Frank [25], the net gain is limited, since other two-dimensional effects come into play for physically thick insulators caused by drain field penetration into the insulator which modulates the channel charge. There is a possible way around this conundrum (**Figure 20**). When the high-permittivity region is confined to the channel area alone, not extending over the heavily doped source and drain, this path is cut off. As noted by Frank [54], this does not change the scale length, since the same equations are solved in the channel cross section; rather, it introduces an extra attenuation in the form of a pre-factor into the equation describing the drain-induced potential along the channel. This pre-factor attenuation is a rapidly increasing function of the permittivity of the dielectric; if it is large enough, it may in itself be sufficient to give the

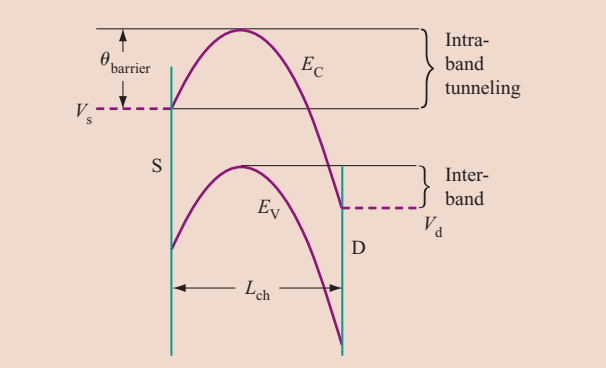


Figure 21

Schematic band diagram along channel of FET, assuming a parabolic profile, showing schematic S–D (intra-band) and band-to-band (inter-band) tunnel-current energy range.

FET the needed electrostatic integrity. In the present case, one may regard the gate insulator as polarizable conduit transferring the potential from the gate to the channel region. Of course, the rest of the FET must be suitably scaled (e.g., the silicon body thickness) so that alternate paths for drain potential feedthrough are minimized. To get a rough idea as to what is involved, take a geometry with insulator height H and gate length L where the insulator has a permittivity ε_G, and the surroundings ε_{sw}. Consider a square of the dielectric closest to the channel (as shown in Figure 20). This couples the dielectric square to the source and drain electrodes with a capacitance $\sim\varepsilon_{sw}$ and to the gate with $\sim\varepsilon_G\ L/H$, per unit width. So, for instance, if a drain vs. gate coupling factor of 0.1 is desired, along with $H/L > 10$, then $\varepsilon_G/\varepsilon_{sw} > 100$ is needed. Therefore, there is motivation to explore materials with very high, and possibly anisotropic, permittivity.

The introduction of this new paradigm removes the gate insulator from the channel-length scaling considerations. For a double-gated FET (and even more so for the surround-gate FET), we are left with a minimum gate length comparable to ($\sim$1.5x greater than) the body thickness. Since FETs with a body thickness <1 nm have been demonstrated, does this mean that a channel length <1.5 nm is feasible? First, we must distinguish between channel length and gate length. Gate length may approach zero, as the vacuum-triode exemplifies, while maintaining basic functionality, if the potential is allowed to drop in the external source/drain regions. For instance, Likharev's simulated 4-nm-gate-length FET [55] actually had a channel length closer to 7 nm when depletion into the contacts was considered. This depletion effect probably characterizes all experimental devices with gate lengths of less than 10 nm.

One could therefore remove the limit on gate length as well. The arbitrary short gate length, for constant channel length, would not result in significantly improved performance because the source-to-drain transit time would still depend on the total distance. In general, such solutions have degraded performance because of their higher resistance. The limit on channel length would be determined by direct drain-to-source tunneling. A simple estimate of this limit is obtained by approximating the potential along the channel in the "off" state by a parabola (**Figure 21**). Such a smooth curve is expected in ultrashort devices, where abrupt transitions are not easy to achieve. The height of the barrier, V_b, determines the "off" current, and the parabola is terminated at "metallic" source and drain contacts where the conduction band is assumed to be pinned at the source and drain potentials. Using Likharev's conjecture [55], we assume that the leakage current is dominated by intra-band tunneling when

$$\frac{kT}{e} \le \frac{\eta}{\pi}\sqrt{\frac{a}{2me}}, \tag{4}$$

where a is the curvature of the parabola, related to the geometry through the source/drain distance L_{ch}, barrier height $\theta_{barrier}$, and drain voltage V_d; m is the tunneling mass, e the electric charge, and kT the thermal voltage. [Tunneling actually dominates before Equation (4) becomes valid due to the $e\theta_{barrier}/kT$ ratio between the tunneling and thermionic emission pre-factors.] Assuming a value of $0.19m_0$ for m, 0.2 eV for $\theta_{barrier}$, and 0.5 V for V_d requires $L_{ch} > 7$ nm to suppress tunneling. Why is this so much larger than Zhirnov's limit [56]? First, the assumption for m (Zhirnov assumed $m = m_0$); next, our assumption of the equality of tunneling and thermionic emission currents, whereas Zhirnov assumed an almost transparent barrier [56]; and finally the use of a parabolic rather than a square barrier. Using a square barrier would reduce our limit to $\sim$4 nm.

We see that limits due to source/drain tunneling, which are not even a factor in current designs, will ultimately limit FET scaling even before electrostatic limits are reached. Power dissipation does not limit scaling on the device level. This is partially because local heat removal from a microscopic part of the device into the 3D surroundings is relatively efficient, and partly because devices operate at a low duty factor, so that local temperature rise is not excessive [57].

A limit that is not often considered, but is intimately related to power dissipation, is band-to-band tunneling. This problem was considered by Solomon [58], primarily for FETs on bulk silicon substrates, but even for SOI devices with ultrathin bodies (UTSOI) and silicon wire devices, this tunneling can be important. As illustrated in Figure 21, we see that there can be band-to-band overlap

when $V_{ds} > E_g - \theta_{barrier}$, where V_{ds} is the drain-to-source voltage, E_g the bandgap, and $\theta_{barrier}$ the barrier height of the turned-off device. For a low-standby-power FET, $\theta_{barrier}$ has to be large in order to limit the leakage current so that the condition of Equation (4) is mostly met. As shown in [58], band-to-band tunneling becomes large when the tunneling distance is ~4 nm. This is typically about 1/3 of the channel length, limiting the channel length in this example to above 12 nm. It is shown in [58] that channel lengths greater than 20 nm are needed for the ultralow-power options of the ITRS roadmap. Since any sharpening of the potential profile (for instance, for a structure with unnecessarily strong electrostatic confinement) would reduce the band-to-band tunneling distance, the design of the "ultimate" FET will involve a delicate tradeoff.

The use of strain to enhance mobilities (Section 4) may enhance band-to-band tunneling. The bandgap of silicon is very sensitive to strain because of the X-valley symmetry of the conduction band; therefore, strain of either sign always decreases the bandgap. The interaction between strain and band-to-band tunneling leakage must therefore be carefully monitored.

New materials

As analyzed in the past [59, 60], the high-mobility III–V materials do not confer an advantage at the end of the scaling path, since the isotropic, low-mass Γ valley does not provide a strong electron confinement and, furthermore, the light electron mass promotes source–drain tunneling. Furthermore, the smooth hetero-barriers, responsible for the spectacular transport, are of low height and do not scale well. The single-transport valley also does not provide the large charge densities needed for these small devices [61].

Materials such as Ge can provide larger carrier densities, but the low bandgap is a considerable disadvantage unless quantum confinement can be used to increase the bandgap in structures of practical dimensions.

How close can we get under current technology assumptions?

The limit of ~7 nm S/D spacing is attainable under current technological assumptions, but not without risk and great effort. Below we discuss the design choices that have to be made, and in the next section we ask whether it is worth it in terms of improved device performance.

While attaining electrostatic integrity is not a limit, it is nonetheless not easy to do, although several paths may be used to achieve it. Electrostatically confined structures also show quantum confinement effects. These can be beneficial (increasing the bandgap) or deleterious (reducing carrier mobility and increasing the sensitivity of the threshold voltage to thickness variations). An apparent advantage for a particular structure, such as a silicon nanowire, in providing superior electrostatic scaling properties may in fact be limited in its scaling potential by the enhanced quantum confinement effects. Wang et al. considered these tradeoffs in the design of planar and wirelike devices [62]. In most cases a wire with a circular cross section showed overall superior scaling properties; however, for a silicon (100) sheet, for electrons, the anisotropic effective mass tensor, with the heavy out-of-plane and light in-plane mass, showed properties superior to those of the wire, as shown in **Figure 22**.

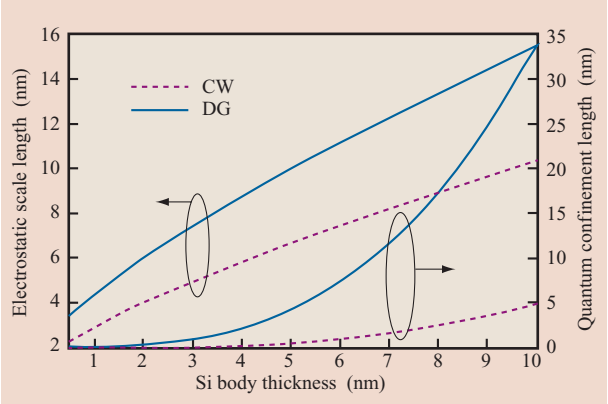


Figure 22

Comparison of electrostatic confinement with quantum confinement for circular wire structures and planar double-gate structures for electrons in a <100> transport direction. From [63], with permission.

Practically, then, what structures can be used and how closely can they approach the limit? Structures based on arrays of silicon wires or strips appear to be the most practical, especially those employing near-planar geometries such as the Tri-Gate, where the advantages of planar geometry are maintained while improved electrostatic control is achieved through the semi-wrapped-around geometry (Section 4). Scaling from designs currently on the drawing board suggests that a FET of 8 nm channel length could be fabricated using silicon strips of 3×10 nm^2 cross section, a gate of ~4 nm length, and an insulator with a relative permittivity of ~100 and a thickness of ~1 nm. To approach the scaling limit without incurring excessive band-to-band tunneling current, voltages must be reduced to well below 1 V. Such devices would have metal rather than the currently used polysilicon gates, since confinement is controlled by geometry rather than doping, and the workfunction of metal gates would thus be a better match to device requirements (Section 4).

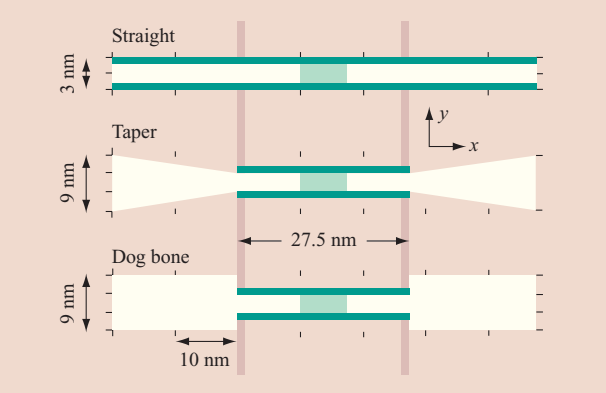


Figure 23

Access geometry for optimal device performance. From [65], with permission.

While ballistic transport has often been touted as the silver bullet to improve performance, as pointed out by Solomon and Laux [61], ballistic transport is actually a limitation to the velocity of the electron caused by its mass being accelerated through a fixed potential. Once ballistic transport sets in, the current will be independent of channel length and the mobility will appear to be inversely proportional to channel length. Ballistic transport is difficult to attain in modern scaled silicon FETs [61, 63], where mobilities are low.

Has the limit already been reached?

As voltages and lengths are reduced toward their "ultimate" limits, leakage currents will tend to increase exponentially. The converse of this is that performance (delay, power) increases only logarithmically with leakage current. We see this in the ubiquitous log-linear I_{off} vs. I_{on} or I_{ddq} vs. delay used to characterize the present generation of devices, whereas a decade ago such curves were rarely seen. Such a tradeoff indicates that a limit has already been reached, since major changes in the operating environment such as temperature or heat-sinking capability, or changes in material properties, result in only modest improvements in performance. Stepwise changes in device structure, probably at great cost, will likewise lead to only stepwise improvements in performance rather than the initiation of a new trend.

Provided we can build a device at the fundamental scaling limit, what advantages would this device bring?

While performance, on average, has increased for each generation of devices, the leading candidates often perform poorly compared with the more optimized and evolved counterparts of the previous generation. Thus, measures of current per unit width or ring-oscillator delay vs. gate length would be maximized, today, at values much larger than the minimum gate lengths on record (~30 nm vs. ~6 nm). It is therefore not appropriate to use current trends as predictors of future performance, especially when the structure of the device is changing so rapidly to meet end-of-scaling challenges.

Modeling of ultimately scaled FETs gives some indication as to their performance potential, although results differ greatly depending on the assumptions used. For instance, Laux et al., using a 2D quantum-mechanical model and assuming ballistic transport, predict currents around 18 A/cm for a supply voltage of 0.4 V [64], whereas, under roughly the same conditions of voltage and leakage currents, Venugopal et al. predict currents of 21 A/cm [65]. If scattering is assumed in [65], the current is reduced by approximately 50%. The same reduction is obtained if a realistic series resistance is included in the calculation.

In both of the papers cited above, the shape and nature of the contacts play a major role in determining the current. For instance, when various geometry contacts (straight, taper, and dog bone) were compared [65] as shown in **Figure 23**, the straight contact gave the largest current, even though it had the smallest cross section. However, when scattering was introduced [65], these differences were minimized. The issue of series resistance is not easily solved; it becomes a fundamental issue as carrier concentrations in the FET channel approach the concentrations in the contacts. For instance, in [64] electron densities in the FET channel were approaching 10^{20} cm^{-2} compared with maximum attainable doping densities in silicon of $\sim 3 \times 10^{20}$ cm^{-3}. Also, to contact the channel directly with a butted metal contact would require an impractically low metal–semiconductor interface resistance of 3×10^{-10} Ω-cm. Thus, the issue of spreading the current into the contact coupled with matching the wavefunctions of the channel to the contacts has to be solved.

The other issue for performance is reduction of capacitance. As discussed in [61], channel capacitances are dominated by degeneracy effects in the quantum limit. Fortunately, for silicon the degeneracy capacitance is very large, 8×10^6 F/cm^2 for the doubly degenerate low-mass sub-bands on (100) silicon. Assuming typical parasitic capacitances of ~0.2 aF/nm at each gate edge means that these capacitances will dominate only at gate lengths below ~0.5 nm. This means that scaling can continue, using suitable gate dielectrics, down to the tunneling limit without being dominated by parasitic capacitance. Performance estimates of extremely scaled devices [64, 65], including a hypothetical device with a high-permittivity dielectric, compared with today's (45-nm ITRS node) base case are compared in **Table 2**. For

the ultimate device, ballistic transport is assumed with capacitance equal to half the degeneracy capacitance, and a series resistance corresponding to the taper in Figure 23 of 13.3 Ω-μm. We can therefore expect at most an additional factor of ~10 performance increase for future silicon FETs.

6. Conclusions

In the previous sections we have shown that short-channel behavior is the limiting factor for device scaling in a power-constrained environment if dielectric scaling halts at current levels of about 1 nm equivalent oxide thickness. In the absence of dielectric scaling, the effective current I_{eff} is not significantly increased with gate-length scaling. The introduction of a metal gate helps to boost the current drive at the cost of higher gate capacitance, which compensates in part for the impact of higher drive current in gate-loaded circuits. Gate-length reduction through continued dielectric scaling, together with a near-band-edge workfunction metal gate, is beneficial for performance. There is a remedy, of course, if technology provides other solutions to enhance drive current. We have discussed the influence of enhanced mobility and have shown that in order to continue performance enhancement, the mobility increase required must go faster than gate-length scaling. To maintain the benefit of strong mobility enhancement, one has to understand the enhancement mechanisms with respect to process flow and device layout, since these can cause a large variation in the strain field seen by the devices. High-mobility substrate options offer the hope of less sensitivity to process and layout details.

We have not included in our investigation the impact of junction engineering and well design, which provide additional levers to modulate short-channel effects. It is clear that series resistance has a considerable effect on drive current [66]. Improved series resistance can be achieved with higher doping activation and improved silicide/diffusion resistance. The specific contact resistance for the silicide/diffusion interface is around 10^{-7} Ω-cm^2 for currently used silicides [67] and achievable doping levels. This gives a transition length of approximately 100 nm. Once the contact length is much shorter, the contact resistance increases dramatically and degrades the drive current. To maintain density shrinkage, a lower silicide/diffusion is required to obtain a smaller transition length. Junction optimization to reduce series resistance is a problem common to all device structures. We have discussed partially depleted SOI, fully depleted SOI on thin body, and several versions of multiple-gate structures. Although fully depleted SOI on thin body has, in principle, a junction capacitance advantage compared with partially depleted SOI, it also has increased source-to-drain capacitance and a higher overlap capacitance due to the raised source process that is required to enable silicide formation. This reduces the ac advantage of fully depleted SOI compared with partially depleted SOI. However, fully depleted SOI on thin body (T_{si} ~ 15 nm) would enable more flexibility in process options such as shallow halo implant or a reduced gate stack height, to enable shallow source/drain implants. It also provides a natural path for a planar back-gated device. Extremely thin silicon body devices can greatly extend the choice of possible workfunctions if the gate-length-to-body-thickness ratio is chosen appropriately. To achieve sufficient short-channel control, the ratio should be larger than 3 to 4 for single-gated, fully depleted, doping-controlled devices.

Table 2 Device on-current for ultimate MOSFET. Device on-currents are taken from literature and normalized to the same gate over drive. C_p is peripheral capacitance. For performance estimate, $4C_p$ is used to account for the Miller effect in ring delay.

	L_g (nm)	V_{dd} (V)	I_{on} (A/cm)	C_g (aF/nm)	$4C_p$ (aF/nm)	*Rel. perf.*
Base SG	25	1	12	0.54	0.8	1.0
Laux DG	7.5	0.4	18	0.32	0.8	2.6
Venugopal DG	10	0.4	21	0.43	0.8	2.9
Venugopal DG (R_s/scatter)	10	0.4	11	0.43	0.8	1.5
"Ultimate" DG	4	0.4	50	0.17	0.8	7.9

Finally, we have made an attempt to predict the performance of the ultimate FET. We have shown that the limiting factor of scaling is the intra-band source-to-drain tunneling and have given an estimate that is more conservative than earlier ones. Our assumption is that due to the finite curvature of the potential well, the channel length is limited to about 7 nm, at which point tunneling will start to dominate over the thermal injection across the barrier. We also argued that the finite thickness of a dielectric with extremely high permittivity will not limit scaling. Provided that we can use all the tricks in the book, we find that the ultimate FET might give one order of performance gain compared with current technology. A realistic value will possibly be smaller, since parasitic elements will diminish the performance of the intrinsic device.

Even if we could build the ultimate device, we would still have to consider its manufacturability and whether we will have a wire and contact technology to connect these devices to take advantage of their performance. Engineering ingenuity and persistence will solve many

of the problems. Finally, the end of scaling will not be decided by what is possible, but by what is affordable. Therefore, the *end of scaling question* will appear in yet another light.

References

1. B. Hoeneisen and C. A. Mead, "Fundamental Limitations in Microelectronics—I. MOS Technology," *Solid State Electron.* **15**, 819–829 (1972).
2. J. T. Wallmark, "Fundamental Physical Limitations in Integrated Electronic Circuits," *Inst. Phys. Conf. Ser.*, No. 25, pp. 133–167 (1975).
3. C. Hu, "Gate Oxide Scaling Limits and Projections," *IEDM Tech. Digest*, pp. 96–99 (1996).
4. J. H. Stathis and D. J. DiMaria, "Reliability Projection for Ultra-Thin Oxides at Low Voltage," *IEDM Tech. Digest*, pp. 167–170 (1998).
5. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, S. Rishton, and E. Ganin, "High Transconductance and Velocity Overshoot in NMOS Devices at the 0.1 mm Gate Length Level," *IEEE Electron Device Lett.* **EDL-9**, p. 464 (1988).
6. R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. LeBlanc, "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits* **SC-9**, 256–268 (1974).
7. Y. Taur, D. Buchanan, W. Chen, D. Frank, K. Ismail, S.-H. Lo, G. Sai-Halasz, R. Viswanathan, H.-J. C. Wann, S. Wind, and H.-S. Wong, "CMOS Scaling into the Nanometer Regime," *Proc. IEEE* **85**, 486–504 (1997).
8. E. P. Gusev, V. Narayanan, and M. M. Frank, "Advanced High-κ Dielectric Stacks with PolySi and Metal Gates: Recent Progress and Current Challenges," *IBM J. Res. & Dev.* **50**, No. 4/5, 387–410 (2006, this issue).
9. 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.* **22**, 95–97 (2001).
10. M. V. Fischetti and S. E. Laux, "Long-Range Coulomb Interactions in Small Si Devices. Part I: Performance and Reliability," *J. Appl. Phys.* **89**, No. 2, 1205–1231 (2001).
11. G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-*K* Gate Dielectrics: Current Status and Materials Properties Considerations," *J. Appl. Phys.* **89**, 5243 (2001).
12. M. V. Fischetti, D. A. Neumayer, and E. A. Cartier, "Reduction of the Electron Mobility in High-*k* MOS Systems Caused by Remote Optical," *High-k Gate Dielectrics*, Michael Houssa, Ed., Institute of Physics, 2004, pp. 397–430.
13. K. Rim, J. L. Hoyt, and J. F. Gibbons, "Transconductance Enhancement in Deep Submicron Strained-Si 12-MOSFETs," *IEDM Tech. Digest*, pp. 707–710 (1998).
14. S. Thompson, N. Anand, M. Armstrong, C. Auth, B. Arcot, M. Alavi, P. Bai, J. Bielefeld, R. Bigwood, J. Brandenburg, M. Buehler, S. Cea, V. Chikarmane, C. Choi, R. Frankovic, T. Ghani, G. Glass, W. Han, T. Hoffmann, M. Hussein, P. Jacob, A. Jain, C. Jan, S. Joshi, C. Kenyon, J. Klaus, S. Klopcic, J. Luce, Z. Ma, B. Mcintyre, K. Mistry, A. Murthy, P. Nguyen, H. Pearson, T. Sandford, R. Schweinfurth, R. Shaheed, S. Sivakumar, M. Taylor, B. Tufts, C. Wallace, P. Wang, C. Weber, and M. Bohr, "A 90 nm Logic Technology Featuring 50nm Strained Silicon Channel Transistors, 7 layers of Cu Interconnects, Low *k* ILD, and 1 μm^2 SRAM Cell," *IEDM Tech. Digest*, pp. 61–64 (2002).
15. H. S. Yang, R. Malik, S. Narasimha, Y. Li, R. Divakaruni, P. Agnello, S. Allen, A. Antreasyan, J. C. Arnold, K. Bandy, M. Belyansky, A. Bonnoit, G. Bronner, V. Chan, X. Chen, Z. Chen, D. Chidambarrao, A. Chou, W. Clark, S. W. Crowder, B. Engel, H. Harifuchi, S. F. Huang, R. Jagannathan, F. F. Jamin, Y. Kohyama, H. Kuroda, C. W. Lai, H. K. Lee, W.-H. Lee, E. H. Lim, W. Lai, A. Mallikarjunan, K. Matsumoto, A. McKnight, J. Nayak, H. Y. Ng, S. Panda, R. Rengarajan, M. Steigerwalt, S. Subbanna, K. Subramanian, J. Sudijono, G. Sudo, S.-P. Sun, B. Tessier, Y. Toyoshima, P. Tran, R. Wise, R. Wong, I. Y. Yang, C. H. Wann, L. T. Su, M. Horstmann, Th. Feudel, A. Wei, K. Frohberg, G. Burbach, M. Gerhardt, M. Lenski, R. Stephan, K. Wieczorek, M. Schaller, H. Salz, J. Hohage, H. Ruelke, J. Klais, P. Huebler, S. Luning, R. van Bentum, G. Grasshoff, C. Schwan, E. Ehrichs, S. Goad, J. Buller, S. Krishnan, D. Greenlaw, M. Raab, and N. Kepler, "Dual Stress Liner for High Performance Sub-45nm Gate Length SOI CMOS Manufacturing," *IEDM Tech. Digest*, pp. 1075–1078 (2004).
16. M. Yang, M. Ieong, L. Shi, K. Chan, V. Chan, A. Chou, E. Gusev, K. Jenkins, D. Boyd, Y. Ninomiya, D. Pendleton, Y. Surpris, D. Heenan, J. Ott, K. Guarini, C. D'Emic, M. Cobb, P. Mooney, B. To, N. Rovedo, J. Benedict, R. Mo, and H. Ng, "High Performance CMOS Fabricated on Hybrid Substrate with Different Crystal Orientations," *IEDM Tech. Digest*, pp. 453–456 (2003).
17. D. Burger and J. R. Goodman, "Billion-Transistor Architectures: There and Back Again," *IEEE Computer*, p. 22 (2004).
18. R. C. Chu, R. E. Simons, M. J. Ellsworth, R. R. Schmidt, and V. Cozzolino, "Review of Cooling Technologies for Computer Products," *IEEE Trans. Device & Mater. Reliabil.* **4**, 568 (2004).
19. G. Baccarani, M. R. Wordeman, and R. H. Dennard, "Generalized Scaling Theory and Its Application to a 1/4 Micrometer MOSFET Design," *IEEE Trans. Electron Devices* **ED-31**, 452 (1984).
20. D. J. Frank, W. Haensch, G. Shahidi, and O. H. Dokumaci, "Optimizing CMOS Technology for Maximum Performance," *IBM J. Res. & Dev.* **50**, No. 4/5, 419–431 (2006, this issue).
21. C.-T. Chuang, K. Bernstein, R. V. Joshi, R. Puri, K. Kim, E. J. Nowak, T. Ludwig, and I. Aller, "Scaling Planar Silicon Devices," *IEEE Circuits & Devices Mag.* **20**, 6–19 (2004).
22. E. J. Nowak, "Maintaining the Benefits of CMOS Scaling When Scaling Bogs Down," *IBM J. Res. & Dev.* **46**, No. 2/3, 169–180 (2002).
23. M. H. Na, E. J. Nowak, W. Haensch, and J. Cai, "The Effective Drive Current in CMOS Inverters," *IEDM Tech. Digest*, pp. 121–124 (2002).
24. P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S.-H. Lee, N. Lindert, M. Liu, Z. Ma, T. Marieb, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber, B. Woolery, A. Yeoh, K. Zhang, and M. Bohr, "A 65 nm Logic Technology Featuring 35 nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-*k* ILD and 0.57 μm^2 SRAM Cell," *IEDM Tech. Digest*, pp. 657–660 (2004).
25. D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, "Device Scaling Limits of Si MOSFETs and Their Application Dependencies," *Proc. IEEE* **89**, 259–288 (2001).
26. M. Ieong, R. Logan, and J. Slinkman, "Efficient Quantum Correction Model for Multi-Dimensional CMOS Simulations," *SISPAD'98 Tech. Digest*, pp. 129–132 (1998).
27. M. B. Ketchen and M. Bhushan, "Product-Representative 'At Speed' Test Structures for CMOS Characterization," *IBM J. Res. & Dev.* **50**, No. 4/5, 451–468 (2006, this issue).
28. R. Dennard, J. Cai, and A. Kumar, "A Perspective on Today's Scaling Challenges and Possible Future Directions," presented at the Conference on the ULtimate Integration of Silicon (ULIS), Grenoble, France, 2006.
29. H.-S. P. Wong, D. J. Frank, and P. M. 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*, pp. 407–410 (1998).

30. D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo Simulation of a 30 nm Dual-Gate MOSFET: How Far Can Si Go?," *IEDM Tech. Digest*, p. 553 (1992).
31. D. Hisamoto, T. Kaga, Y. Kawamoto, and E. Takeda, "A Fully Depleted Lean-Channel Transistor (DELTA)—A Novel Vertical Ultra Thin SOI MOSFET," *IEDM Tech. Digest*, pp. 833–836 (1989).
32. K. Sunouchi, H. Takato, N. Okabe, T. Yamada, T. Ozaki, S. Inoue, K. Hashimoto, K. Hieda, A. Nitayama, F. Horiguchi, and F. Masuoka, "A Surrounding Gate Transistor Cell for 64/256 Mbit DRAMs," *IEDM Tech. Digest*, pp. 23–36 (1989).
33. X. Huang, W.-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y.-K. Choi, K. Asano, V. Subramanian, T.-J. King, J. Bokor, and C. Hu, "Sub 50-nm FinFET: PMOS," *IEDM Tech. Digest*, pp. 67–70 (1999).
34. J. Kedzierski, D. M. Fried, E. J. Nowak, T. Kanarsky, J. H. Rankin, H. Hanafi, W. Natzle, D. Boyd, Y. Zhang, R. A. Roy, J. Newbury, C. Yu, Q. Yang, P. Saunders, C. P. Willets, A. Johnson, S. P. Cole, H. E. Young, N. Carpenter, D. Rakowski, B. A. Rainey, P. E. Cottrell, M. Ieong, and H.-S. P. Wong, "High-Performance Symmetric-Gate and CMOS-Compatible V_t Asymmetric-Gate FinFET Devices," *IEDM Tech. Digest*, pp. 437–440 (2001).
35. T. Park, S. Choi, D. H. Lee, J. R. Yoo, B. C. Lee, J. Y. Kim, C. G. Lee, K. K. Chi, S. H. Hong, S. J. Hynn, Y. G. Shin, J. N. Han, I. S. Park, U. I. Chung, J. T. Moon, E. Yoon, and J. H. Lee, "Fabrication of Body-Tied FinFETs (Omega MOSFETs) Using Bulk Si Wafers," *Symp. VLSI Technol.*, pp. 135–136 (2003).
36. B. Doyle, B. Boyanov, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, R. Rios, and R. Chau, "Tri-Gate Fully-Depleted CMOS Transistors: Fabrication, Design and Layout," *Symp. VLSI Technol.*, pp. 133–134 (2003).
37. K. W. Guarini, P. M. Solomon, Y. Zhang, K. K. Chan, E. C. Jones, G. M. Cohen, A. Krasnoperova, M. Ronay, O. Dokumaci, J. J. Bucchignano, C. Cabral, Jr., C. Lavoie, V. Ku, D. C. Boyd, K. S. Petrarca, I. V. Babich, J. Treichler, P. M. Kozlowski, J. S. Newbury, C. P. D'Emic, R. M. Sicina, and H.-S. Wong, "Triple-Self-Aligned, Planar Double-Gate MOSFETs: Devices and Circuits," *IEDM Tech. Digest*, pp. 425–428 (2001).
38. A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, "Impact of Intrinsic Fluctuations on CMOS SRAM Cell Stability," *IEEE J. Solid-State Circuits* **36**, No. 4, 658–665 (2001).
39. D. J. Frank, Y. Taur, M. Ieong, and H.-S. P. Wong, "Monte Carlo Modeling of Threshold Variation Due to Dopant Fluctuations," *Symp. VLSI Technol.*, pp. 169–170 (1999).
40. S. A. Mujtaba, "Advanced Mobility Models for Design and Simulation of Deep Submicron MOSFETs," Ph.D. dissertation, Stanford University, CA, 1995.
41. K. Uchida, H. Watanabe, A. Kinoshita, J. Koga, T. Numata, and S. Takagi, "Experimental Study on Carrier Transport Mechanism in Ultrathin-Body SOI n- and p-MOSFETs with SOI Thickness Less Than 5 nm," *IEDM Tech. Digest*, pp. 47–50 (2002).
42. D. Esseni, A. Abramo, L. Selmi, and E. Sangiorgi, "Study of Low Field Electron Transport in Ultra-Thin Single and Double Gate SOI MOSFETs," *IEDM Tech. Digest*, pp. 719–722 (2002).
43. H. Sakaki, T. Noda, K. Hirikawa, M. Tanaka, and T. Matsusue, "Interface Roughness Scattering in GaAs/AlAs Quantum Well," *Appl. Phys. Lett.* **51**, No. 23, 1934–1936 (1987).
44. M. V. Fischetti, F. Gamizy, and W. Haensch, "On the Enhanced Electron Mobility in Strained-Silicon Inversion Layers," *J. Appl. Phys.* **92**, No. 12, 7320 (2002).
45. M. V. Fischetti, Z. Ren, P. M. Solomon, M. Yang, and K. Rim, "Six-Band k.p Calculation of the Hole Mobility in Silicon Inversion Layers: Dependence on Surface Orientation, Strain, and Silicon Thickness," *J. Appl. Phys.* **94**, 1079–1095 (2003).
46. C.-H. Chen, T. L. Lee, T. H. Hou, C. L. Chen, C. C. Chen, J. W. Hsu, K. L. Cheng, Y. H. Chiu, H. J. Tao, Y. Jin, C. H. Diaz, S. C. Chen, and M.-S. Liang, "Stress Memorization Technique (SMT) by Selectively Strained Nitride Capping for Sub-65 nm High-Performance Strained-Si Device Application," *Symp. VLSI Technol.*, p. 56 (2004).
47. K. Rim, K. Chan, L. Shi, D. Boyd, J. Ott, N. Klymko, F. Cardone, L. Tai, S. Koester, M. Cobb, D. Canaperi, B. To, E. Duch, I. Babich, R. Carruthers, P. Saunders, G. Walker, Y. Zhang, M. Steen, and M. Ieong, "Fabrication and Mobility Characteristics of Ultra-Thin Strained Si Directly on Insulator (SSDOI) MOSFETs," *IEDM Tech. Digest*, pp. 47–52 (2003).
48. A. Oishi, O. Fujii, T. Yokoyama, K. Ota, T. Sanuki, H. Inokuma, K. Eda, T. Idaka, H. Miyajima, S. Iwasa, H. Yamasaki, K. Oouchi, K. Matsuo, H. Nagano, T. Komoda, Y. Okayama, T. Matsumoto, K. Fukasaku, T. Shimizu, K. Miyano, T. Suzuki, K. Yahashi, A. Horiuchi, Y. Takegawa, K. Saki, S. Mori, K. Ohno, I. Mizushima, M. Saito, M. Iwai, S. Yamada, N. Nagashima, and F. Matsuoka, "High Performance CMOSFET Technology for 45 nm Generation and Scalability of Stress-Induced Mobility Enhancement Technique," *IEDM Tech. Digest*, pp. 239–242 (2005).
49. T. Sanuki, A. Oishi, Y. Morimasa, S. Aota, T. Kinoshita, R. Hasumi, Y. Takegawa, K. Isobe, H. Yoshimura, M. Iwai, K. Sunouchi, and T. Noguchi, "Scalability of Strained Silicon CMOSFET and High Drive Current Enhancement in the 40 nm Gate Length Technology," *IEDM Tech. Digest*, pp. 65–68 (2003).
50. G. Eneman, P. Verheyen, R. Rooyackers, F. Nouri, L. Washington, R. Degraeve, B. Kaczer, V. Moroz, A. De Keersgieter, R. Schreutelkamp, M. Kawaguchi, Y. Kim, A. Samoilov, L. Smith, P. P. Absil, K. DeMeyer, M. Jurczak, and S. Biesemans, "Layout Impact on the Performance of a Locally Strained PMOSFET," *Symp. VLSI Technol.*, pp. 22–23 (2005).
51. J. Cai, K. Rim, A. Bryant, K. Jenkins, C. Ouyang, D. Singh, Z. Ren, K. Lee, H. Yin, J. Hergenrother, T. Kanarsky, A. Kumar, X. Wang, S. Bedell, A. Reznicek, H. Hovel, D. Sadana, D. Uriarte, R. Mitchell, J. Ott, D. Mocuta, P. O'Neil, A. Mocuta, E. Leobandung, R. Miller, W. Haensch, and M. Leong, "Performance Comparison and Channel Length Scaling of Strained Si FETs on SiGe-on-Insulator (SGOI)," *IEDM Tech. Digest*, pp. 165–168 (2004).
52. J. R. Brews, "Physics of the MOS Transistor," *Appl. Solid State Sci.*, Suppl. 2A, pp. 1–120 (1981).
53. K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, "Scaling Theory for Double-Gate SOI MOSFET's," *IEEE Trans. Electron Devices* **40**, 2326–2329 (1993).
54. D. J. Frank and H.-S. Wong, "Analysis of the Design Space Available for High-k Gate Dielectrics in Nanoscale MOSFETs," *Superlat. & Microstruct. (UK)* **28**, 485–491 (2000).
55. F. G. Pikus and K. K. Likharev, "Nanoscale Field Effect Transistors—An Ultimate Size Analysis," *Appl. Phys. Lett.* **71**, 3661 (1997).
56. V. V. Zhirnov, R. K. Cavin III, J. A. Hutchby, and G. I. Bourianoff, "Limits to Binary Logic Switch Scaling—A Gedanken Model," *Proc. IEEE* **91**, 1934 (2003).
57. H.-S. Wong, D. J. Frank, P. M. Solomon, C. H. J. Wann, and J. Welser, "Nanoscale CMOS," *Proc. IEEE* **89**, 259 (2001).
58. P. M. Solomon, D. J. Frank, J. Jopling, C. D'Emic, O. Dokumaci, P. Ronsheim, and W. E. Haensch, "Universal Tunneling Behavior in Technologically Relevant PN Junctions," *J. Appl. Phys.* **95**, No. 10, 5800–5812 (2004).
59. P. M. Solomon, "Physics at the Limit of VLSI Scaling," *AIP Conf. Proc.*, No. 122, *The Physics of VLSI*, J. C. Knights, Ed., 1984, p. 172.
60. M. V. Fischetti and S. E. Laux, "Monte Carlo Simulation of Electron Transport in Technologically Significant

Semiconductors of the Diamond and Zinc-Blende Structures. Part II: Submicron MOSFETs," *IEEE Trans. Electron Devices* **ED-38**, 650 (1991).
61. P. M. Solomon and S. E. Laux, "The Ballistic FET: Design, Capacitance and Speed Limit," *IEDM Tech. Digest*, pp. 511–514 (2001).
62. J. Wang, P. Solomon, and M. Lundstrom, "A General Approach for the Performance Assessment of Nanoscale Silicon Field Effect Transistors," *IEEE Trans. Electron Devices* **51**, 1366 (2004).
63. 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.* **22**, 95 (2001).
64. S. E. Laux, A. Kumar, and M. V. Fischetti, "Analysis of Quantum Ballistic Electron Transport in Ultrasmall Silicon Devices Including Space-Charge and Geometric Effects," *J. Appl. Phys.* **95**, 5545 (2004).
65. 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.* **95**, 292 (2004).
66. S. D. Kim, S. Narasimha, and K. Rim, "An Integrated Methodology for Accurate Extraction of S/D Series Resistance Components in Nanoscale MOSFETs," *IEDM Tech. Digest*, pp. 155–158 (2005).
67. M. C. Ozturk, "Source/Drain Junctions and Contacts for 45 nm CMOS and Beyond," *Proceedings of the International Conference on Characterization and Metrology for ULSI Technology*, 2005; see *http://www.eeel.nist.gov/812/conference/2005_presentations.html*.

Received October 1, 2005; accepted for publication April 27, 2006; Internet publication August 3, 2006

Wilfried Haensch *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (whaensch@us.ibm.com)*. In 1981, Dr. Haensch received his Ph.D. degree from the Technical University of Berlin, Germany, in the field of theoretical solid-state physics. In 1984 he joined Siemens Corporate Research in Munich to investigate high-field transport in MOSFET devices, and in 1988 he joined the DRAM development team at the Siemens Research Laboratory to investigate new cell concepts. In 1990, he joined the DRAM alliance between IBM and Siemens to develop quarter-micron 64M DRAM. In this capacity, Dr. Haensch was involved with device characterization of shallow-trench bounded devices and cell-design concerns. In 1996, he moved to a manufacturing facility to build various generations of DRAM. His primary mission was to transfer technologies from development into manufacturing and to guarantee a successful yield ramp of the product. In 2001, Dr. Haensch joined the IBM Thomas J. Watson Research Center to lead a group concerned with novel devices and applications. He is currently responsible for post-45-nm-node device design and its implications for circuit functionality.

Edward J. Nowak *IBM Systems and Technology Group, 1000 River Street, Essex Junction, Vermont 05452 (ejnowak@us.ibm.com)*. Dr. Nowak received his B.S. degree in physics in 1973 from M.I.T., and M.S. and Ph.D. degrees, also in physics, from the University of Maryland in 1975 and 1978, respectively. In 1981, following postdoctoral research at New York University, he joined IBM in Essex Junction, Vermont, to work on DRAM development. Since 1985, Dr. Nowak has worked in high-performance CMOS device design. His current interests include energy-driven device design and FinFET device architectures.

Robert H. Dennard *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (dennard@us.ibm.com)*. Dr. Dennard is an IBM Fellow at the IBM Thomas J. Watson Research Center, where he works on extending scaling toward its fundamental limits and on energy-efficient computing. In 1954 and 1956 he received B.S. and M.S. degrees in electrical engineering from Southern Methodist University. He received the Ph.D. degree from the Carnegie Institute of Technology in 1958 and then joined the IBM Research Division. Dr. Dennard has been involved in microelectronics research and development from its early days. In 1967 he invented the one-transistor dynamic memory cell (DRAM) used in most computers today. With co-workers, he developed the concept of MOSFET scaling in 1972. Dr. Dennard is a Fellow of the IEEE and received their Edison Medal in 2001. He is a member of the National Academy of Engineering and the American Philosophical Society. His honors include the National Medal of Technology presented by President Reagan in 1988 and induction into the National Inventors Hall of Fame in 1997.

P. M. Solomon *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (solomonp@us.ibm.com)*. Dr. Solomon studied electrical engineering as an undergraduate at the University of Cape Town, South Africa, in 1968; he received a Ph.D. degree from the Israel Institute of Technology in 1975. Since joining IBM in 1975, Dr. Solomon has been a Research Staff Member at the Thomas J. Watson Research Center, where he has worked on aspects of MOS, bipolar, and III–V device physics and device design, probing the limits of device scaling. He is a Fellow of the IEEE and the American Physical Society.

Andres Bryant *IBM Systems and Technology Group, 1000 River Street, Essex Junction, Vermont 05452 (bryanta@us.ibm.com)*. Dr. Bryant received his B.S.E.E. degree from the University of Maine in 1982 and his Ph.D. degree in

electrical engineering from Stanford University in 1986, joining IBM in Burlington, Vermont, that same year. His work areas have ranged from surface-acoustic-wave gas sensors and scanning tunneling microscopy to CMOS transistor design. He has also worked on DRAM transistor design and high-performance-logic transistor design. Dr. Bryant's current interests include energy-efficient, ultralow-voltage transistor and circuit design.

Omer H. Dokumaci *IBM Systems and Technology Group, 2070 Route 52, Hopewell Junction, New York 12533 (dokumaci@us.ibm.com)*. Dr. Dokumaci received his Ph.D. degree in electrical engineering from the University of Florida, joining the Process and Device Modeling Group at the IBM facility in Hopewell Junction, New York. Dr. Dokumaci's research has concentrated on modeling and simulation of dopant diffusion and activation, and advanced devices such as FinFETs, ultrathin silicon, metal-gate, back-gate, and ground-plane devices.

Arvind Kumar *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (arvkumar@us.ibm.com)*. Dr. Kumar received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology. After graduation, he pursued postdoctoral studies in mesoscopic shot noise at CEA Saclay, France. In 1996, he joined the IBM Thomas J. Watson Research Center, where his primary interests have been in semiconductor device physics, modeling, and design.

Xinlin Wang *IBM Systems and Technology Group, 2070 Route 52, Hopewell Junction, New York 12533 (xinlinw@us.ibm.com)*. Ms. Wang received her B.S. degree in electronics engineering from Tsinghua University, China, in 1998 and her M.S. degree in electrical engineering from the University of Massachusetts at Amherst in 2001. That same year she joined the Device and Process Modeling group at the IBM Semiconductor Research and Development Center in Hopewell Junction, New York. Her research interests focus on device physics, modeling, and simulation of nanoscale transistors.

Jeffrey B. Johnson *IBM Systems and Technology Group, 1000 River Street, Essex Junction, Vermont 05452 (jbj@us.ibm.com)*. Dr. Johnson received the Ph.D. degree in electrical and computer engineering in 1987 from Carnegie Mellon University. That same year he joined IBM in Essex Junction, Vermont, working in the field of technology simulation, both software development and program application. He has applied process and device simulation to many generations of IBM technology, including DRAM, SOI and bulk CMOS, and SiGe BiCMOS programs. Dr. Johnson is currently a Senior Technical Staff Member in the Technology Simulation and Predictive Modeling Department.

Massimo V. Fischetti *Department of Electrical and Computer Engineering, 201D Marcus Hall, University of Massachusetts, Amherst, Massachusetts 01003 (fischett@ecs.umass.edu)*. Dr. Fischetti received the Ph.D. degree in physics from the University of California at Santa Barbara in 1978. From 1983 to 2004 he was a Research Staff Member at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York, working on the theory of electron transport in semiconductors and insulators and on Monte Carlo simulation of semiconductor devices. In 2005 he joined the Department of Electrical and Computer Engineering at the University of Massachusetts, Amherst. His interests cover electronic properties of semiconductors and insulators and semiclassical and quantum transport.

Errata

In the paper "Silicon CMOS Devices Beyond Scaling" by W. Haensch et al. in the *IBM Journal of Research and Development*, Volume 50, No. 4/5, July/September 2006, the exponent 2 was omitted from the expression fCV^2 in the body and caption of Figure 1. The corrected figure follows.

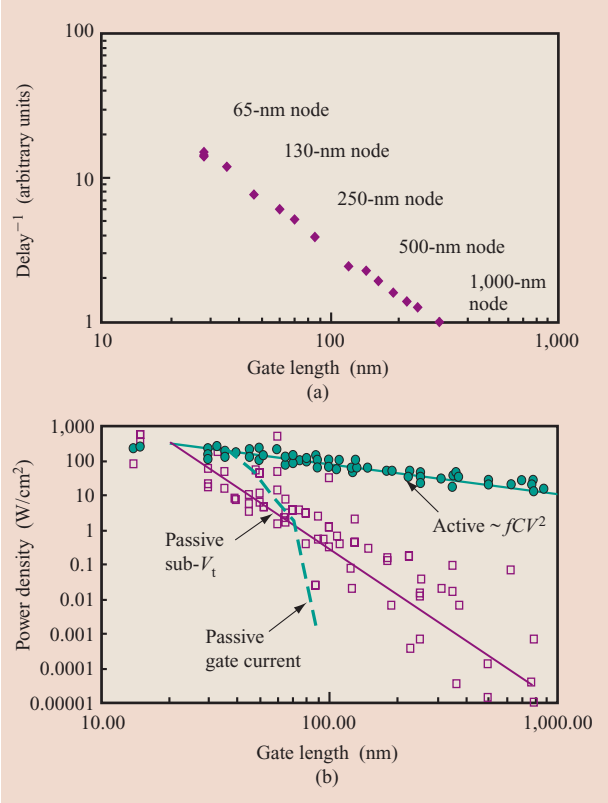


Figure 1

(a) MOSFET performance vs. gate length; normalized MOSFET intrinsic device delay (CV/I_{eff}) vs. gate length. (b) Power density vs. gate length; data collected from literature for active power density and passive power density. Lines are intended to show trend. (fCV^2 = frequency × capacitance × voltage^2.)

In the paper "Continuous MOSFET Performance Increase with Device Scaling: The Role of Strain and Channel Material Innovations" by D. A. Antoniadis et al. in the *IBM Journal of Research and Development*, Volume 50, No. 4/5, July/September 2006, the last term of Equation (5) should be multiplied by v. The corrected equation follows.

$$\begin{aligned} I_{\text{eff}} &= [I_D(V_G = V_{dd}/2, V_D = V_{dd}) \\ &\quad + I_D(V_G = V_{dd}, V_D = V_{dd}/2)]/2 \\ &= [Q_s'(V_G = V_{dd}/2, V_D = V_{dd}) \\ &\quad + Q_s(V_G = V_{dd}, V_D = V_{dd}/2)]v/2 \\ &= C'_{\text{oxinv}} W[(3-\delta)V_{dd}/4 - V_t]v. \end{aligned} \tag{5}$$