Device Scaling Limits of Si MOSFETs and Their Application Dependencies

DAVID J. FRANK, MEMBER, IEEE, ROBERT H. DENNARD, FELLOW, IEEE, EDWARD NOWAK, MEMBER, IEEE, PAUL M. SOLOMON, FELLOW, IEEE, YUAN TAUR, FELLOW, IEEE, AND HON-SUM PHILIP WONG, FELLOW, IEEE

Invited Paper

This paper presents the current state of understanding of the factors that limit the continued scaling of Si complementary metal-oxide-semiconductor (CMOS) technology and provides an analysis of the ways in which application-related considerations enter into the determination of these limits. The physical origins of these limits are primarily in the tunneling currents, which leak through the various barriers in a MOS field-effect transistor (MOSFET) when it becomes very small, and in the thermally generated subthreshold currents. The dependence of these leakages on MOSFET geometry and structure is discussed along with design criteria for minimizing short-channel effects and other issues related to scaling. Scaling limits due to these leakage currents arise from application constraints related to power consumption and circuit functionality. We describe how these constraints work out for some of the most important application classes: dynamic random access memory (DRAM), static random access memory (SRAM), low-power portable devices, and moderate and high-performance CMOS logic. As a summary, we provide a table of our estimates of the scaling limits for various applications and device types. The end result is that there is no single end point for scaling, but that instead there are many end points, each optimally adapted to its particular applications.

Keywords*—CMOS, device design, discrete dopants, double-gate MOSFET, DRAM, high-k dielectrics, high-performance logic, leakage currents, limits, low power, MOSFET, nanotechnology, power density, scale length, scaling, SRAM, tunneling.*

I. INTRODUCTION

In 1930, Lilienfeld [1] patented the basic concept of the field effect transistor (FET) . Thirty years later, in 1960, it was finally reduced to practice in Si–SiO_2 by Kahng and Attala [2]. Since that time, it has been incorporated into integrated circuits and has grown to be the most important device in the electronics industry. Progress in the field for at least the last 25 years has followed an exponential behavior that has come to be known as Moore's Law [3]. Since 1994, the semiconductor industry has been projecting these exponentials into the future to provide technology development targets. The most recent of these projections is the 1999 International Technology Roadmap for Semiconductors (ITRS99) [4]. It contains projections for complementary metal-oxide-sesmiconductor (CMOS) technology out to 2014, including 32-Gb dynamic random access memory (DRAM) entering production and processors with gate lengths down to 20 nm and 2×10^{10} FETs per chip.

But will these exponential projections come to pass or will physical limits make them impossible? Many reviews have been written about the current state and future prospects for Si MOS field-effect transistors (MOSFETs) and CMOSs [5]–[9]. In particular, many different scaling limits for MOSFETs have been proposed and discussed. In this work, we describe the current state of understanding of these scaling limits and seek to advance this state of understanding by addressing the ways in which application requirements must be intertwined with the setting of limits. The result in the end is that there will be no single "end to scaling," but rather, a wide range of limiting FET technologies, each optimally adapted to its applications.

Much of our discussion centers on bulk-like MOSFET scaling, as illustrated in Fig. 1, but this is not intended to exclude other device geometries for MOSFETs. In particular, partially depleted silicon-on-insulator (PD-SOI) MOSFETs are considered to be part of this bulk-like category, since most of the same limits apply to PD-SOIs as to bulk. Consequently, PD-SOI is not explicitly discussed except when there are significant device design differences. At the circuit level, there are, of course, some important features of SOIs, such as the floating body effects, but these are for the most part outside the scope of this paper. The scaling behavior of fully depleted silicon-on-insulator (FD-SOI) MOSFETs depends a great deal on the

Manuscript received March 30, 2000; revised September 26, 2000.

The authors are with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: djf@us.ibm.com; pwong@watson.ibm.com).

Publisher Item Identifier S 0018-9219(01)02066-7.

thickness of the buried oxide. For thick buried oxide, there is no backside screening of the drain potential, resulting in relatively poor scaling characteristics compared to other device types [10]–[13]. Since such devices are not likely to be used at the limits of scaling they are not discussed here. We do, however, discuss the scaling advantages of the more novel double gated type of FD-SOI MOSFETs, wherein both the insulator on the back side of the Si channel layer and the Si layer itself are very thin so that both sides of the channel are gated. There are also in-between FD-SOI MOSFETs with buried oxide thin enough to offer some screening, but not thin enough for use in active switching. These devices are interesting from a circuit point of view since the back gate can be used to dynamically adjust the threshold voltage, but are not discussed here for lack of space.

The outline of the paper is as follows. Section II addresses some of the more fundamental limitations to the continued scaling of MOSFETs that appear to be on the horizon. Based only on these fundamental limits, it may be possible to scale FETs down to very small dimensions, e.g., 10-nm channel length or smaller. Section III describes research results related to this fundamental limit regime: very tiny one-of-a-kind FETs. In the more practical world of manufacturing, however, there are many types of variations and fluctuations that require the design of MOSFETs with tolerances. In Section IV, we look at some of these practical limitations and their consequences for device design. Section V describes how the concepts of the previous sections play out when they are applied to meeting the needs of specific classes of applications. The paper ends in Section VI by summarizing all of the limits into a large table, followed by the conclusion in Section VII.

II. Fundamental Scaling Limits

A. Scaling Theory

For many years now, the shrinking of MOSFETs has been governed by the ideas of scaling [14], [15]. The basic idea is illustrated in Fig. 1: a large FET is scaled down by a factor α to produce a smaller FET with similar behavior. When all of the voltages and dimensions are reduced by the scaling factor α and the doping and charge densities are increased by the same factor, the electric field configuration inside the FET remains the same as it was in the original device. This is called constant field scaling, which results in circuit speed increasing in proportion to the factor α and circuit density increasing as α^2. These scaling relations are shown in the second column of Table 1 along with the scaling behavior of some of the other important physical parameters.

Fig. 2 illustrates the actual past and projected future scaling behavior of several of these parameters versus the channel length [16]. As can be seen, the voltages have not been scaled at the same rate as the length, in violation of the simple scaling rules outlined above. In earlier generations of MOSFETs, this occurred because carrier velocities were increasing with increasing field, yielding higher performance, while deleterious high-field effects were kept in check by the gradually descending voltage. More recently, carrier velocities have become saturated, but voltage scaling has

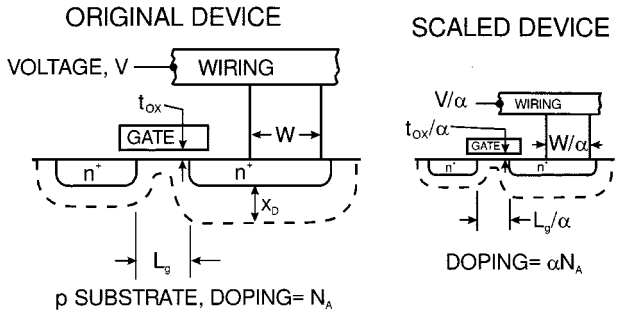


Fig. 1. Schematic illustration of the scaling of Si technology by a factor alpha. Adapted from [5].

Table 1
Technology Scaling Rules for Three Cases

Physical parameter	Constant-Electric Field Scaling Factor	Generalized Scaling Factor	Generalized Selective Scaling Factor
Channel length, Insulator thickness	$1/\alpha$	$1/\alpha$	$1/\alpha_d$
Wiring width, channel width	$1/\alpha$	$1/\alpha$	$1/\alpha_w$
Electric field in device	1	ε	ε
Voltage	$1/\alpha$	ε/α	ε/α_d
On-current per device	$1/\alpha$	ε/α	ε/α_w
Doping	α	$\varepsilon\alpha$	$\varepsilon\alpha_d$
Area	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha_w^2$
Capacitance	$1/\alpha$	$1/\alpha$	$1/\alpha_w$
Gate delay	$1/\alpha$	$1/\alpha$	$1/\alpha_d$
Power dissipation	$1/\alpha^2$	ε^2/α^2	$\varepsilon^2/\alpha_w\alpha_d$
Power density	1	ε^2	$\varepsilon^2\alpha_w/\alpha_d$

α is the dimensional scaling parameter, ε is the electric field scaling parameter, and α_D and α_W are separate dimensional scaling parameters for the selective scaling case. α_D is applied to the device vertical dimensions and gate length, while α_W applies to the device width and the wiring.

been slow because of the nonscaling of the subthreshold slope and the OFF current. To accommodate this trend, more generalized scaling rules have been created, in which the electric field is allowed to increase by a factor ε [17]. Furthermore, the device widths and wiring dimensions have not been scaled as fast as the channel lengths, leading to a further scaling parameter for those dimensions. These generalized rules are also shown in Table 1 and are described in more detail in [5], [9], and [18].

The preceding scaling rules do not tell a designer how short he can make a MOSFET for given doping profiles and layer thicknesses; they only describe how to shrink a known good design. Furthermore, since the built-in potentials are not usually scaled, the rules are inaccurate anyway. To find the minimum gate length at each generation of technology, one must analyze the two-dimensional (2-D) field effects inside the FET. This is often done numerically using complex 2-D simulation tools, but the recent analytic analysis by Frank *et al.* [19] reveals the primary dependencies. Other

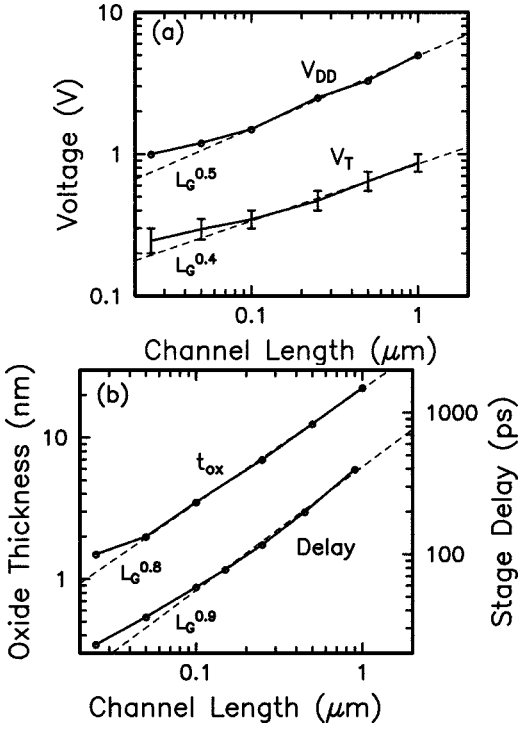


Fig. 2. Past and projected future scaling trends for CMOS logic. (a) Supply voltage and threshold voltage versus channel length. (b) Gate oxide thickness and 2-in NAND delay versus channel length. Adapted from [16].

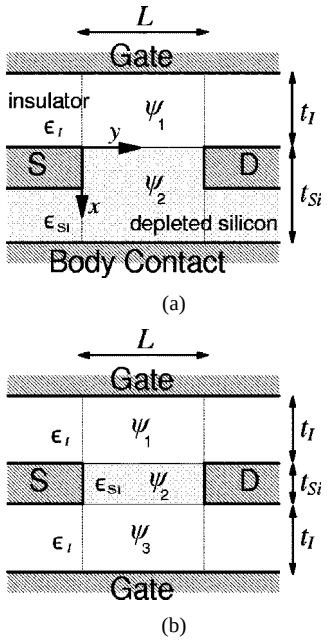


Fig. 3. Idealized schematic cross section diagrams of (a) a bulk MOSFET and (b) a DG-FET, defining the insulator thickness t_I and the depleted Si thickness t_{Si}. ϵ_{Si} is the dielectric constant of silicon and ϵ_I is the dielectric constant of the gate insulator(s). Adapted from [19].

analyses have been made in the past [10], [20], [21], but we prefer this approach because it allows us to treat the high-k dielectric case accurately.

According to this theory, the details of which are summarized in the Appendix, the potential variations in the channel

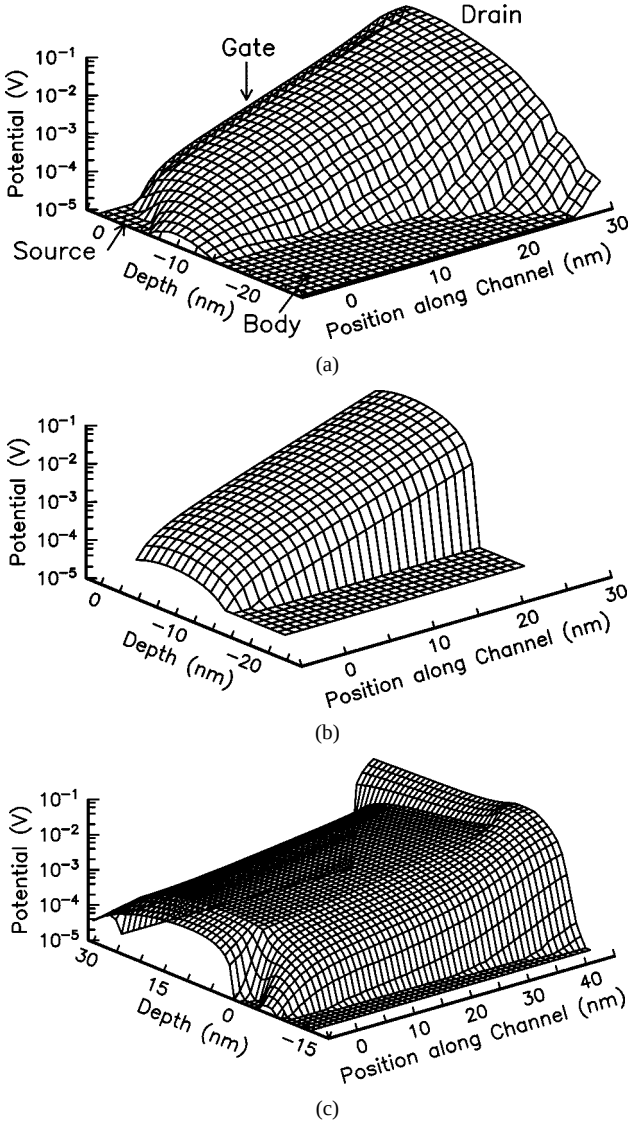


Fig. 4. 2-D potential perturbations in nFETs caused by a 20-mV variation in the drain potential. (a) 2-D numeric simulation for realistic doping profiles. (b) Simple analytic theory using the same conditions as (a). (c) 2-D numeric simulation for a high-k gate insulator ($k = 78$, 30 nm thick) with extreme ground-plane-like doping profiles and shallow source and drain. From [33].

of an idealized MOSFET structure such as that in Fig. 3(a) can be expressed analytically using functions of the form $\sinh(n\pi y/\Lambda_n)\sin(n\pi x/\Lambda_n)$. The full dielectric boundary conditions can be satisfied by matching these functions at the interface, leading to an implicit equation for the scale length Λ_1, which characterizes the lowest order solution

$$0 = \epsilon_{\rm Si}\tan(\pi t_I/\Lambda_1) + \epsilon_I \tan(\pi t_{\rm Si}/\Lambda_1) \qquad (1)$$

where the symbols are defined in Fig. 3. In the most common regime, $t_I/\Lambda_1 \ll 1$ and (1) can be approximately solved as $\Lambda_1 \simeq t_{\rm Si} + (\epsilon_{\rm Si}/\epsilon_I)t_I - (\pi^2/3)(\epsilon_{\rm Si}/\epsilon_I)(\epsilon_{\rm Si}^2/\epsilon_I^2 - 1)(t_I/t_{\rm Si})^2 t_I$. There is also an analogous scale length for the double-gate MOSFET (DG-FET), which is a three-layer structure with a gate and a thin gate insulator on both sides of the channel, as shown schematically in Fig. 3(b). Its equation is given in the Appendix. Fig. 4(a) and (b) shows

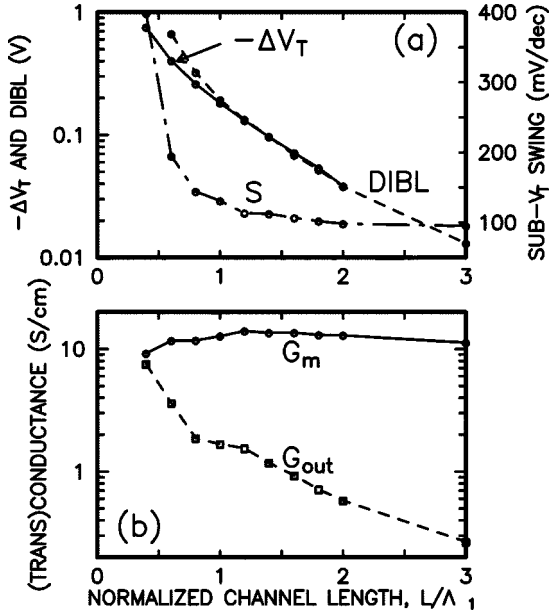


Fig. 5. Plots of (a) ΔV_T, DIBL, and inverse subthreshold slope and (b) transconductance (G_m) and output conductance (G_{out}), all versus the L/Λ_1 ratio, showing the dependence of short channel effects on channel length. Based on 2-D FIELDAY simulations of idealized FET structures with $\Lambda_1 = 13.6$ nm ($t_{ox} = 1.5$ nm, $t_{Si} = 10$ nm). ΔV_T is determined at $V_{DS} = 0.05$ V, DIBL is defined as $V_T(V_{DS} = 0.05) - V_T(V_{DS} = 1.0)$, the transconductance is measured at $V_{DS} = 1.0$ V, $V_G = V_T(V_{DS} = 0.05) + 0.5$ V and the output conductance is measured at the same V_G, and $V_{DS} = 0.75$ V.

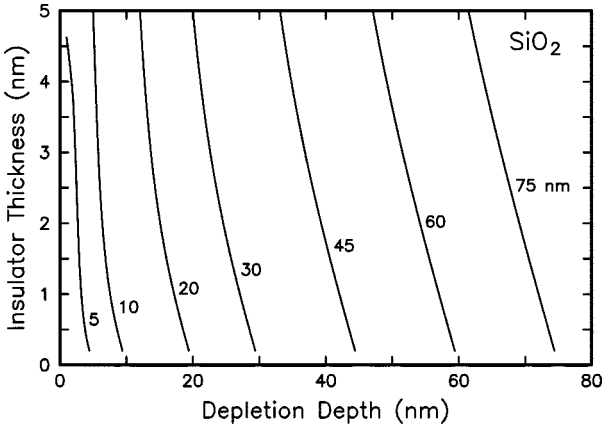


Fig. 6. Plot of constant Λ_1 contours versus t_I and t_{Si} for $\epsilon_{Si}/\epsilon_I = 3$. From Frank *et al.* [19].

a comparison between the numerically calculated 2-D potential change in a conventional MOSFET due to a change in drain voltage and the first-order analytic approximation. Clearly, the simple approximation accurately captures the functional form of the potential variation along the channel, where it is most important. The only substantial difference is in the deep depletion under the drain, but this does not significantly influence the subthreshold behavior.

For this lowest order solution, the source–drain component of the potential in the center of the channel varies as $(b_{21}+c_{21})\sinh(\pi L/2\Lambda_1)/\sinh(\pi L/\Lambda_1)$, where b_{21} and c_{21} are bias dependent. Since this gives a length dependence of $\sim\exp(-\pi L/2\Lambda_1)$, the L/Λ_1 ratio is a fundamental measure of the quality of the FET. For $L/\Lambda_1 \gg 1$, the FET will behave nearly ideally according to the one-dimensional (1-D) gradual channel approximation, but for small L/Λ_1 there will be strong 2-D effects, including drain-induced barrier lowering (DIBL), high-output conductance, and V_T rolloff. The dependence of these effects on L/Λ_1 is shown in Fig. 5 for a particular case using 2-D numerical simulations of FETs with idealized doping profiles like those in Fig. 3(a). Evidently, $L/\Lambda_1 \gtrsim 0.4$ is a fundamental limit on MOSFET aspect ratio for this idealized design since voltage gain, given by G_m/G_{out}, needs to be greater than one for CMOS logic [22].

This scale length thus transforms the minimum gate length question into a question of maximum tolerable 2-D effects. From an idealized theoretical point of view, these effects can be large and L/Λ_1 down to around one can be considered, as discussed in Section III. For current manufacturing tolerance ratios and circuit design techniques, however, it appears that the minimum practical worst case short FETs have $L/\Lambda_1 \simeq 1.2$ (for DIBL < 150 mV), so that the minimum practical nominal design point is around $L_{nom}/\Lambda_1 \simeq 1.5$, allowing for $\pm 20\%$ gate length variation. This is only possible because the V_T rolloff in bulk MOSFETs is partially compensated by lateral doping nonuniformity (e.g., halo doping, see Section IV-A). For uniform lateral doping (e.g., an undoped DG-FET), it is probably necessary to have $L_{nom}/\Lambda_1 \simeq 1.7$ just to keep ΔV_T below $\sim$100 mV, although the exact ratio probably depends on the desired V_T. The bulk limit can be seen, for example, in recent manufacturing technology [23] in which the minimum gate length (100 nm) FETs have DIBL of $\sim$120 mV (at $V_{DS} = 1$ V) and G_m/G_{out} of $\sim$10, which correspond well to the $L/\Lambda_1 = 1.2$ point in Fig. 5. High V_{DD} threshold rolloff at this point would be unacceptably high, except that it is largely canceled out by careful halo doping.

As a specific example of this scale length, Fig. 6 shows the numerically evaluated dependence of Λ_1 from (1) on t_I and t_{Si} for the Si–SiO_2 system. The simple linear approximation $\Lambda_1 = t_{Si} + (\epsilon_{Si}/\epsilon_I)t_I$ corresponds well to the $\Lambda_1 = 75$ nm case in Fig. 6, but note that the slope of the contours increases dramatically for shorter scale lengths, indicating a significant departure from this approximate solution. This increased slope is beneficial to highly scaled FETs since it implies that the penalty for using insufficiently scaled oxide thickness is less than might have been expected.

By their nature, none of these scaling rules contain in their formulation any limit on how far they can be applied. The limits enter due to physical phenomena that are not included in the scaling. The physical dimensions are limited by quantum mechanical tunneling currents that pass through the various barriers in the MOSFET when they are sufficiently thin, degrading the device's behavior. Voltage scaling is limited on several fronts. The built-in junction voltages are set by the 1.1-eV bandgap of Si which does not scale. Consequently, as the applied voltages are scaled down toward 1 V, the internal fields do not automatically scale as desired. A similar difficulty occurs in trying to scale

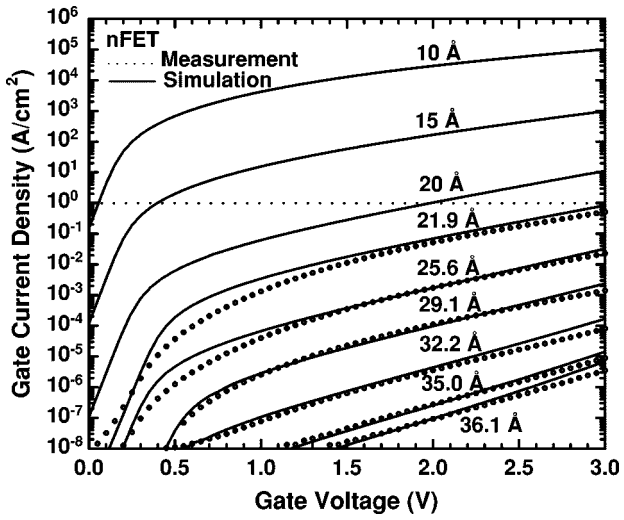


Fig. 7. Calculated (lines) and experimental (dots) results for tunnel currents from inversion layers through thin oxides. Adapted from Lo *et al.* [24].

the threshold voltage V_T, which is tied to the nonscaling behavior of the subthreshold slope and its influence on the OFF current. At very low values, the supply voltage is also fundamentally limited by the need for sufficient gain to provide logic functionality. These and other issues associated with scaling are examined in more detail below. Currently, the scaling of physical dimensions is also limited in a practical sense by the discreteness of dopants since present manufacturing techniques do not control the exact placement of dopant atoms. Consequently, since very small device volumes contain only a small number of dopants, large statistical variations become likely. Although single devices can be built, large functional circuits may be unmanufacturable by present techniques. This effect is discussed in Section IV-C.

B. Tunneling Limits

Tunneling current through the gate insulator is one of the most constraining limits to scaling. For SiO_2, the conventional gate insulator, this leakage (see Fig. 7) exceeds the requirements of some applications (e.g., DRAM) already at 2.5–3 nm, even though high-performance logic technology is currently pushing 2-nm oxide thickness [23] to achieve the desired performance. According to Fig. 7 [24], 2-nm SiO_2 will have a leakage current of about 0.1 A/cm^2 at 1.2 V. For conventional designs this will only contribute a few milliwatts to the overall chip dissipation, which is only problematic for very low power applications, but is indicative of where things are headed. Several writers suggest that the upper limit of acceptable gate leakage is in the 1–10-A/cm^2 range [25]–[27] or even 100 A/cm^2 [28], although if one assumes more aggressively that up to 10% of the total power dissipation could be due to gate leakage (see Section V-A), then it may be possible to tolerate leakage $\sim$1000 A/cm^2 in very high performance chips and even higher densities in small areas (unless reliability problems prevent it, see Section V-A). Either way, the minimum pure SiO_2 gate insulator thickness for high-performance applications is in the 1.0–1.5-nm range, which should be reached in one or two generations. Lower power applications require thicker minimum oxide thickness and are already near their limits.

What can be done to circumvent this limit? There are at least three paths of attack, all of which may be useful. The first approach is to stop scaling the oxide, but attempt to continue scaling the rest of the FET in such a manner as to compensate for the thicker oxide. A related approach is to change the device structure in such a way that the MOSFET can be scaled further, even with the relatively thicker oxide. DG-FETs are an example of this approach. The third approach is to try to change the gate insulator to another material such that the effective capacitive thickness can be reduced without increasing the tunneling current.

The first approach has two aspects: 1) one can reduce the depletion depth (or the Si layer thickness for DG-FETs) as far as possible, to minimize Λ_1 without further thinning of the oxide and 2) one can seek ways to reduce the minimum acceptable L/Λ_1, such as improving the halo doping. The depletion depth can be reduced by increasing the doping and/or by forward biasing the body-source junction, but this has two drawbacks: body leakage currents and a degraded ideality factor. The leakage currents are due to forward body-source junction current and band-to-band tunneling between the body and drain, which is described below. The ideality factor η is the reciprocal of the rate of change of the channel surface potential as a function of V_G (in the subthreshold regime) and is approximately equal to $1 + \epsilon_{Si} t_I / \epsilon_I t_{Si}$ for bulk MOSFETs. It enters into the subthreshold slope, thus impacting the OFF current. Applications that can be refrigerated may particularly benefit in this regime, since forward junction current and degraded subthreshold slope can both be ameliorated by running at low operating temperature.

As mentioned briefly in the preceding section, a laterally nonuniform doping distribution can at least partially compensate for the V_T rolloff that occurs for $L/\Lambda_1 < 2$. Of all the 2-D effects, V_T rolloff has the worst effect on circuit margins, so this compensation is very important and enables worst case L/Λ_1 to be reduced from $\sim$1.5 for no compensation to $\sim$1.2 for current generation halo doping. Halo doping achieves this lateral nonuniformity by angling in shallow body-type doping from the source and drain ends of the FET with the gate as a mask creating a "halo" around the source and drain. For shorter FETs, these halo profiles work to create a higher average doping in the channel than is seen by a longer channel FET, thus tending to raise the V_T in opposition to short-channel effects that are lowering it. Such halos are used to achieve the 25-nm bulk CMOS design described in Section IV-A. The other 2-D effects, however, are not compensated and for L/Λ_1 much below 1.2, device performance becomes severely impacted anyway.

The second approach involves changing the device structure to one in which the gate essentially surrounds the channel. The most investigated form is the DG-FET in which there are a gate and a thin gate insulator on both sides of the channel, as shown schematically in Fig. 3(b). This geometry has been shown to have better scaling properties than the conventional bulk MOSFET [11], [12], [29] at least

for room temperature operation and is described in some detail in Section IV-B. The three-dimensional (3-D) version of these devices in which the channel is a thin post and the gate wraps around it cylindrically has the best electrostatic scaling properties of all and has been investigated by several groups [30], [31], but may prove to be impractical because of the high quantization energy levels for such a channel. (The lowest quantum energy level of the confined channel adds to the classical V_T of the MOSFET, creating an additional V_T control issue [6].) The primary advantages of these alternate device structures are a better ideality factor, near unity, and the possibility of thinner Si channels than would be possible in bulk devices except at very low temperature. It is not yet known to what extent V_T rolloff can be compensated in these structures, although it seems that at least in principle, it may be possible to do so. For planar forms of the device, one could implant halo doping profiles into the channel although this would be subject to more fluctuations than for bulk devices because the volume available for such doping is smaller due to the thinness of the channel. Lateral variations in the gate workfunction might also be possible [32].

Finally, there is much work aimed at reducing the gate tunneling problem by changing to a higher permittivity (k) gate insulator. This is largely a materials problem since its success depends upon achieving high layer uniformity, integration with other Si processes, minimal/controlled reactions with Si and the gate electrode, and low fixed-charge, defect, and trap densities in the insulator and at the interface between the insulator and the Si substrate. Interface chemistry might also necessitate the use of metallic gate electrodes in which case metals must be found with workfunctions near those of n- and p-poly-Si to achieve low V_Ts. If a suitable insulator can be found, it would be characterized by three thicknesses: its physical thickness t_I, its equivalent oxide tunneling thickness t_{oxTeq}, and its equivalent oxide capacitive thickness t_{oxCeq}. Although t_I would be larger than the (application dependent) minimum SiO_2 film thickness for most high-k dielectrics, the goal is to find an insulator with the property that when its t_{oxTeq} is equal to the minimum SiO_2 thickness, its t_{oxCeq} is significantly less than the minimum SiO_2 thickness. This would enable further scaling since when the gate insulator permittivity varies, at least initially, all of the other device dimensions and voltages can be scaled in keeping with t_{oxCeq} rather than the physical thickness t_I (since this maintains the scaling of charge density).

There are, however, some constraints on high-k insulators. The scale length theory of Section II-A shows that the physical thickness of the high-k insulator becomes important as k increases, increasing the scale length and the drain potential penetration under the gate [19], [33]. This is illustrated in Fig. 4(c), which shows the potential perturbation in the channel of a MOSFET with $k = 78$ gate insulator and the same t_{oxCeq} as in (a) and (b). Note that the potential falls much more slowly in (c) than in (a) and (b), even though the channel is longer in (c). A more detailed analysis of Λ_1 from (1) shows that the physical insulator thickness should always be less than the Si depletion depth under the channel since otherwise the scale length will actually increase with

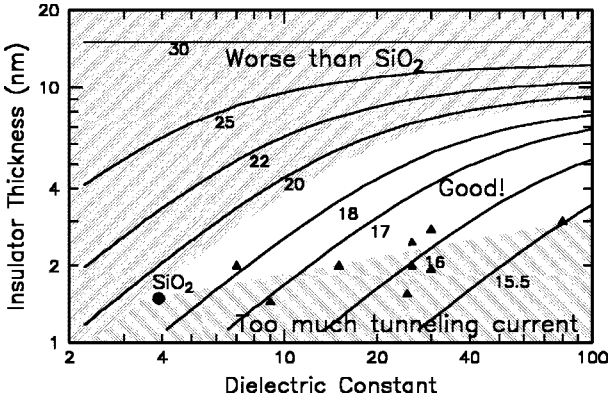


Fig. 8. Contours of constant scale length versus dielectric constant and insulator thickness, showing the useful design space for high-k gate dielectrics. Data points are rough estimates of the tunneling constraints for various high-k insulators. Depletion depth is 15 nm here. Useful design space will shrink with decreasing depletion depth. From [33].

increasing k [19]. This may be thought of as a case of "majority rule": when there is more insulator than Si, one approaches the situation in which $\Lambda_1 = t_I + (\epsilon_I/\epsilon_{Si})t_{Si}$, i.e., the Si gets converted to equivalent insulator thickness rather than the insulator being converted to equivalent Si.

The overall implications of the scale length considerations on high-k dielectrics are illustrated in Fig. 8, which indicates the regime in which these dielectrics can usefully contribute to further scaling of Si MOSFETs. The contours of constant Λ_1 are equivalent to contours of minimum gate length since minimum gate length is proportional to Λ_1. Since it is undesirable to retreat from scaling and be forced to make larger FETs, the upper region that corresponds to larger minimum FETs than can be achieved with Si–SiO_2 is blocked out. The lower region is blocked out by the approximate tunneling leakage limits of high-k materials and reflects the empirical observation that insulator bandgaps tends to decrease with increasing k. Only the unhatched region is usefully available for high-k improvements to scaling. Based on the ratio between Λ_1 for SiO_2 (19 nm) and the best accessible Λ_1 at high k (15.5 nm), it appears that high-k materials can offer about one additional generation of gate length scaling at fixed t_{Si}, but probably not more.

The gate insulator is not the only barrier through which tunneling currents may flow in very small MOSFETs. The body-to-drain junction can also experience tunneling currents if the field is high enough. Fig. 9 shows the field dependence of such band-to-band tunneling currents. Since the cross-sectional area of the highest field body-to-drain junction region is ~1/3 that of the gate insulator, it may be possible to tolerate higher tunneling current density, perhaps up to 3000 A/cm^2 for aggressive high-end applications. According to Fig. 9, this puts the scaling limit for body-to-drain electric field at ~2.7 MV/cm, which corresponds to peak body doping around 3×10^{19} cm^{-3} for bulk MOSFETs, depending on bias and doping gradients, for a minimum depletion depth t_{Si} of 8–13 nm. For low-power applications, the limit is likely to be below 1 A/cm^2 or 1.7 MV/cm and $0.8–1.2\times10^{19}$ cm^{-3} body doping for a minimum t_{Si} of

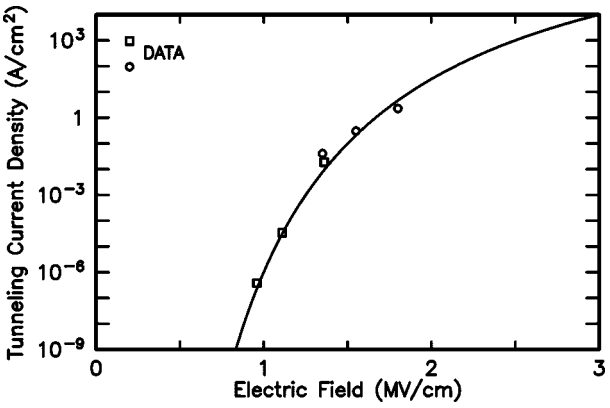


Fig. 9. Plot of band-to-band pn junction tunneling current versus electric field for 1 V reverse bias. Adapted from [27].

13–15 nm. In SOI MOSFETs with floating bodies, including DG-FETs, this tunneling current is potentially more problematic because, in addition to creating dissipation, it can charge up the floating body, lowering the effective threshold voltage. This body charging reaches steady state when the body voltage is low enough that the thermionic current into the source balances out the tunneling injection from the drain. Since the tunneling current depends strongly on the drain-to-source voltage, so does the body bias, which can create very high output conductance. For DG-FETs, the situation is not quite so bad: the rate at which carriers that have tunneled from the drain into the body can thermionically exit the body into the source is approximately the same as the rate at which carriers are thermally injected into the channel from the source. Therefore, as long as the drain-to-body tunneling current does not exceed the subthreshold channel current, the device should behave reasonably well. For more conventional SOI FETs, however, the barrier for carriers to leave the body can be much higher than the channel barrier and therefore very little drain-to-body tunneling can be tolerated before the $\mathrm{V_T}$ is shifted significantly. The only apparent way around this floating body problem for FD-SOI is to lower the supply voltage so that there is no direct tunneling path available between drain and body. This imposes an approximate constraint on $\mathrm{V_{DD}}$ for conventional FD-SOI of the form $\mathrm{V_{DD}} \leq E_G/\mathrm{e} - \mathrm{V_T}/\eta$. For PD-SOI, there is also a second solution available: use a body contact.

It must be pointed out that these tunneling estimates could easily be too optimistic because the currents in Fig. 9 are for ideal band-to-band tunneling. Such tunneling can be greatly enhanced by deep traps in the junction, resulting in much higher junction leakage currents that would depend on the statistical distribution of deep traps in the junctions. This problem particularly impacts DRAM retention time distributions [34], [35].

For *very* small MOSFETs, direct subthreshold source-to-drain tunneling through the potential barrier below the gate is another possible source of leakage current. This effect has been reported in electrically variable shallow junction MOSFETs (EJ-MOSFETs) operating at 77 K with physical gate lengths of 8 nm [36] and is expected to become important at room temperature for channel lengths around 10 nm (see Section III). It currently appears that scaling for most applications will stop due to minimum insulator thickness problems before this source-to-drain tunneling limit is reached.

One last tunneling-related constraint on scaling MOSFETs is tied to the need for FETs in most applications to provide greater than unity power gain, not to mention voltage gain. For sufficiently leaky gate insulators, the power required to drive the input leakage current could exceed the power available at the output especially if the output conductance is high. This would turn the FET into an attenuator rather than an amplifier and so represents perhaps the ultimate limit on thinning the gate insulator. For practical very large scale integration (VLSI) applications, however, power density problems are likely to limit scaling long before this limit is reached.

C. Voltage Limits

The most conspicuous nonscaling voltage in the conventional Si MOSFET is the Si bandgap potential $E_G/\mathrm{e} = 1.1\,\mathrm{V}$ (where e is the elementary charge), which can only be changed significantly by changing the semiconductor itself. This nonscaling behavior does not actually limit operating voltage, but it does complicate device design. In traditional circuit design, the body is tied to the source supply voltage and, consequently, as the supply voltage is scaled down into the 1-V range, the effect of the bandgap potential is increasing. The primary effect is to increase the junction fields and/or depletion depths in the FET above what they would be for ideal scaling. For the body-to-source and -drain junctions, the higher field necessitates higher junction doping, but the nonscaled E_G tends to suppress the band-to-band tunneling compared to what it would be if the bandgap were scaled. For the channel depletion region, the increasing field perpendicular to the oxide interface confines channel carriers closer to the interface, reduces their mobility, increases their quantum confinement energy, and increases gate depletion. Since these effects tend to increase the threshold voltage, they make it very hard to lower $\mathrm{V_T}$ to the levels needed for high-performance applications. The first step in achieving a lower $\mathrm{V_T}$ and channel surface field while still getting a scaled shallow depletion depth is to use retrograde doping profiles with low doping at the surface and high doping near the desired depletion depth [6]. If this does not lower $\mathrm{V_T}$ sufficiently for some applications, one could consider very shallow counterdoping of the surface of the retrograde-doped channel to further lower the $\mathrm{V_T}$ without significantly increasing the depletion depth. Alternately, most of these E_G scaling problems can be addressed by forward biasing the body relative to the source [37] in a manner which in effect scales E_G. The problems with forward biasing the body include the need to generate and distribute more supply voltages and the forward-biased diode current, which would add to dissipation. Since the latter problem might be solved by low-temperature operation, forward body bias may indeed be a viable solution for high-performance computing applications and is discussed more extensively in Section IV-A. Note that PD-SOI MOSFETs tend to acquire

moderate forward body bias automatically in the process of equalizing the impact ionization and tunneling currents entering and leaving the body.

The biggest limit to scaling $\mathrm{V_T}$ is that the OFF current I_{off} of the FET is constrained by application considerations and $I_{\mathrm{off}} \cong I_{\mathrm{VT}} 10^{-\mathrm{V_T/S}}$, where S is the inverse subthreshold slope and I_{VT} is the current at which $\mathrm{V_T}$ is defined. Since $\mathrm{S} \simeq (\ln 10)\eta kT/e$, where η is the ideality, k is Boltzmann's constant, and T is the temperature, the only way to scale $\mathrm{V_T}$ without also changing I_{off} is to scale T. For high-end applications, this is beginning to happen to some extent, but for many applications (e.g., cell phones), significant cooling is not an option. For these low-to-moderate power applications, the maximum dissipation-limited active-mode I_{off} may range between 10^{-7} and 10^{-4} A/cm, resulting in minimum $\mathrm{V_T}$s varying between 0.54 and 0.27 V, respectively. These thresholds assume $I_{\mathrm{VT}} = 0.1$ A/cm and $\mathrm{S} = 90$ mV/decade. Very high-performance circuits might tolerate thresholds near 100 mV (by this $\mathrm{V_T}$ definition). Note that these are worst case thresholds at high $\mathrm{V_{DD}}$. Nominal threshold voltages must be set higher to allow for manufacturing tolerances. As noted before, double-gate structures generally have smaller inverse subthreshold slope, perhaps 70 mV/decade at room temperature, allowing the threshold and, hence, the supply voltage to be scaled further.

For very low-power applications, there is an interest in reducing the supply voltage as far as possible as a way of reducing the power by trading off performance [38]. From a fundamental point of view, in binary digital logic, the minimum permissible logic swing is the smallest swing that is still large enough to maintain two distinct logic states and it was shown long ago that this level is around $4\eta kT$ [22]. This estimate can be refined by considering the self consistency required for a combinatorial logic gate. In this regime, each logic state is identified with a relatively small range of voltages, either high or low. Self consistency means that a combinatorial logic gate with any possible combination of inputs taken from the logic state ranges will always produce an output state that lies in one of the logic state ranges.

Conventional combinatorial logic circuits are built using series and/or parallel combinations of input devices, one (or two for CMOS) device(s) for each input. To find the fundamental limits, imagine circuits in which the usual sources of logic state degradation (noise) are absent: there are no voltage drops in the wiring, no capacitive coupling between wires, no process-induced parameter variations among the devices, no variations in the supply voltages, no extrinsic resistances, and no thermal noise. The nonlinearity of the active devices serves to compress the variety of input logic states into just two output states, but since the devices are not "infinitely nonlinear," a finite voltage range is required to achieve adequate compression. As an example, Fig. 10 illustrates this for a simple CMOS four-input NAND gate, where (a) shows the bias conditions which lead to the upper and lower logic state ranges. In Fig. 10(b), the logic swing is "large"; the "eye" diagram shows a small amount of noise margin between the earliest switching gate with only one input changing and the latest switching gate with all of its inputs changing and the output state ranges are isolated and self consistent, even though the range of input states does create some spread. This logic swing is above the minimum limit. When the logic swing is reduced too far [see Fig. 10(d)], the earliest and latest curves no longer cross, indicating that there is no self-consistent solution for $\mathrm{V_{L2}}$ and $\mathrm{V_{H2}}$ (as defined in the figure). The lack of a self-consistent state means that operating a long chain of such logic gates can result in the loss of the logic signal (Fig. 11). Fig. 10(c) shows the minimum logic swing condition: the earliest and latest curves are exactly tangent at their intersection points (and the noise margin is reduced to zero).

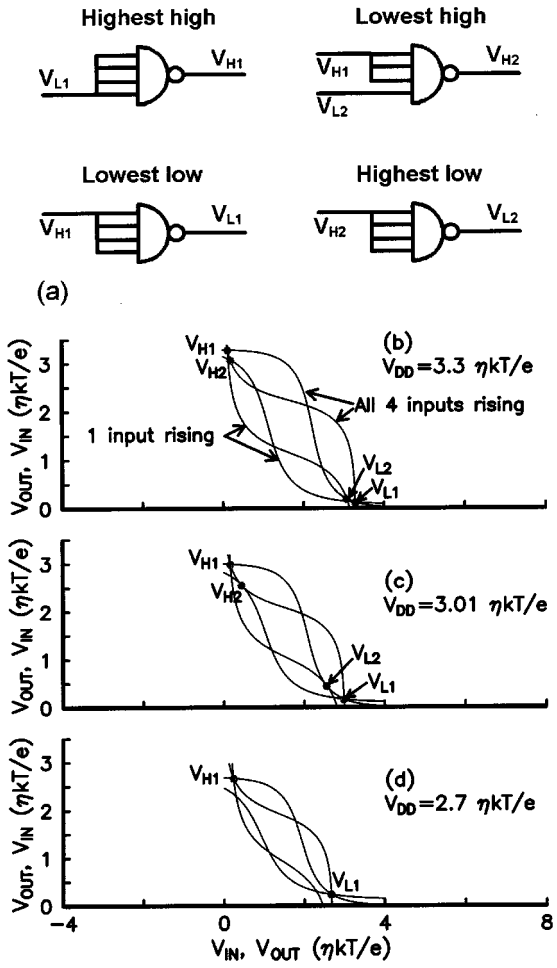


Fig. 10. Illustration of minimum swing determination using ideal four-input CMOS NAND gates. (a) Set of schematics defining conditions for best- and worst case logic outputs. (b)–(d) Transfer characteristics in the form of "eye" diagrams. Output voltage versus input voltage and input voltage versus output voltage for the cases when only one input is switching and when all four inputs are switching. Inputs that are not switching are held at $\mathrm{V_{H1}}$. (b) Logic swing is above the minimum. Self-consistent low output states are between $\mathrm{V_{L1}}$ and $\mathrm{V_{L2}}$ and self-consistent high output states are between $\mathrm{V_{H1}}$ and $\mathrm{V_{H2}}$. (c) Logic swing is exactly at minimum. (d) Logic swing is below the minimum; there are no fully self-consistent output logic states.

Using this type of minimum logic swing condition, analogous curves can be found for other logic families and fan-ins and analytic calculations and simple circuit simulations can be carried out to determine the minimum logic swing or supply voltage; some results for MOSFETs are given in

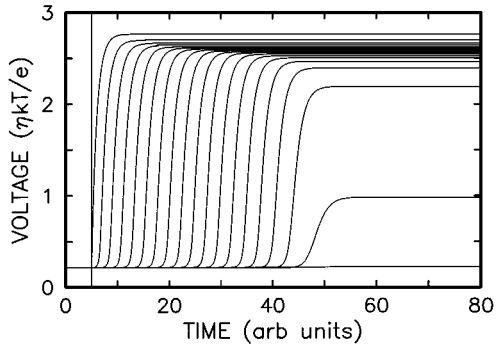


Fig. 11. Output signal versus time for every second gate in a series chain of subthreshold CMOS four-input NAND gates when operated with logic swing of $2.99\eta kT$, which is below the minimum ($3.01\eta kT$). Logic gates are configured in worst case fashion: low-going stages receive all four inputs from the previous stage, while high-going stages receive only one input from the previous stage, having the other inputs tied high. Note that for this case the logic signal is lost after 35 stages.

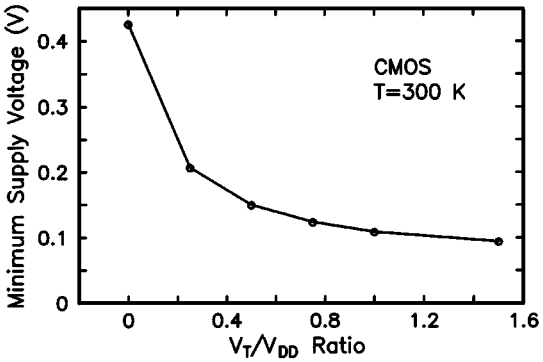


Fig. 12. Minimum supply voltage versus $\mathrm{V_T}$ as a fraction of $\mathrm{V_{DD}}$ for four-input CMOS NAND gates. These calculations use a constant mobility Brews model [97] to simulate 1.0-μm channel length surface channel FETs with thin oxide (4.5 nm) and realistic source–drain contact resistances. Threshold voltage here is defined by the extrapolation to zero of the source–drain conductance for very low drain voltages.

Table 2
Minimum Self-Consistent Supply Voltage for Fixed Fan-In Logic Gates for Several Circuits and Conditions

	Minimum Supply Voltage			
Circuit family	FI=2	FI=4	FI=8	Units
CMOS NAND well below V_T	2.27	3.01	3.72	ηkT/e
CMOS NAND V_T=V_{DD}/4, T=300 K	144.	207.	274.	mV
Resistor Pull-up nFET NOR, well below V_T	4.22	5.2	6.08	ηkT/e

Minimum design points assume that the device sizes and/or bias voltages have been optimized to center the input/output curves. For random logic, the minimum logic swing would be determined by the average or typical fan-in rather than by the worst case, since high fan-in gates would be buffered by lower fan-in circuits.

Table 2. These limits vary roughly as $(\eta kT/e)\ln(\mathrm{FI})$ for conventional devices in their exponential regime, where FI is the fan-in. Note that the lowest voltage FET results are achieved by using the FETs in their subthreshold regime, where they present their maximum exponential nonlinearity. To achieve smaller minimum logic swing, one would need devices with stronger nonlinearities since the greater the nonlinearity, the smaller the voltage range required for logic state compression. Using MOSFETs in the conventional above-threshold manner decreases their overall nonlinearity and increases the required minimum supply voltage, as shown in Fig. 12, from 75 mV (FI = 4) for pure subthreshold CMOS to 207 mV for $\mathrm{V_T} = \mathrm{V_{DD}}/4$ at 300 K. Adding more contact resistance may also increase the minimum supply voltage since it decreases overall nonlinearity[39]. Another consideration is that transient simulations show increased timing variability (dependence on input state) near the minimum logic swing limit because of the asymmetric switching.

The introduction of realistic nonidealities such as noise, tolerances, and short-channel MOSFET behavior into the above analysis will create a statistical spread of scenarios requiring higher voltages to guarantee logic state consistency. On the other hand, from a theoretical but impractical point of view, one could buffer the output of every multiple input logic gate with a chain of inverters, effectively increasing the gain and decreasing the minimum required supply voltage somewhat, although the power consumed by the buffers seems likely to eliminate any real advantage from this approach.

D. Resistance Issues

It is implicitly assumed in the scaling theories that the parasitic resistance in series with the intrinsic MOSFET is either negligible or scalable along with the channel resistance. Otherwise, the performance gains derived from scaling are quickly lost. For example, recent experimental work on 20-nm gate length MOSFETs reported current levels much below today's optimized 100-nm devices because of excessive series resistance [40].

In spite of this case, series resistance is not expected to impose a fundamental limit on CMOS scaling. Technological advances, e.g., self-aligned silicide for contact resistance reduction and rapid thermal annealing for abrupt source–drain formation, allow today's state-of-the-art high-performance bulk nMOSFETs to achieve a series resistance below 100 $\Omega\,\mu$m [23]. This is less than 10% of the effective device resistance $(\mathrm{V_{DD}} - \mathrm{V_T})/\mathrm{I_{on}} \simeq 1000\ \Omega\,\mu$m. Ultimately, the intrinsic device resistance of an ideal ballistic MOSFET approaches $(C_{\mathrm{ox}}v_s)^{-1}$, where C_{ox} is the effective gate capacitance per unit area including quantum effects and v_s is the thermal injection velocity at the source [41]. For a physical or equivalent $\mathrm{t_{ox}}$ of 1.0 nm, the limiting intrinsic device resistance is about 500 $\Omega\,\mu$m. Even without further reduction in series resistance below currently achieved values, no serious performance degradation is expected.

For bulk CMOSs, there is a tendency for the series resistance to increase as the junction depth is scaled down for

shorter channel devices. But this is unlikely to pose a fundamental problem as it can be dealt with by structural solutions such as raised source–drain using selective epi. Furthermore, it is shown in Section IV-A that for an optimized halo design, strict junction depth scaling is not required for short channel control. This can be understood from the principle of the scale length model in which the source–drain depth only enters the preexponential factor of the threshold voltage rolloff.

Particular attention is needed to avoid high series resistance in SOI and/or DG-FETs that use thin silicon films. Ideally, the source and drain regions should fan out to a much thicker film for reduction of both the electrical and the thermal resistance.

III. Ultimate MOSFETs

As indicated in the discussion about scale length and minimum channel length in Section II-A, the primary constraint on shrinking channel length is the coupling between 2-D short channel effects and tolerances. When 2-D effects become large at very short channel length, random variations in gate length, dopant positions, and other structural parameters cause very large changes in device characteristics. If one is only interested in a single FET or if one assumes that ways can eventually be found to reduce process variations to insignificance and to place dopants exactly, then tolerances are not an issue and one can design and build extremely short gate-length MOSFETs. These can be very useful for exploring the physics of small FETs even if they do not reflect manufacturable processes.

The smallest reported experimental FETs are 8-nm EJ-MOSFETs made by Kawaura *et al.* [36]. As shown in the cross-sectional diagram in Fig. 13(a), these electrically variable shallow junction nMOSFETs use a second gate over the top of the first gate to induce inversion layers in the source and drain regions. Such inversion layers are much shallower than the usual implanted source–drain extensions, which reduces to a minimum the influence of the drain on the channel of the FET. The lower gates were patterned by e-beam lithography and lateral etching to achieve a minimum physical gate length of 8 nm. This FET has substrate doping of 2×10^{18} cm^{-3}, a depletion depth of 25–30 nm, $t_{ox} = 5$ nm, and consequently $\Lambda_1 \sim 40$ nm (see Fig. 6). If the effective channel length is of the same order as the gate length, this MOSFET has $L/\Lambda_1 \sim 0.2$, which is extremely small and seems unlikely since its V_T shift, DIBL, and g_m/g_{out} ratio are all consistent with $L_{eff}/\Lambda_1 \sim 0.5$–$0.6$, judging by Fig. 5. By this analysis, it appears likely that this FET has an effective channel length of $\sim$22 nm with the extra $\sim$14 nm due to fringe screening of the upper gate field by the lower gate. Nevertheless, for transport measurements, this FET is very interesting. It appears to be so short that it shows evidence of direct tunneling between source and drain through the channel barrier. This is demonstrated by the inverse subthreshold slope measurements versus temperature shown in Fig. 13(b). The saturation of the inverse slope at low temperatures for the shortest FETs is consistent with the idea that the current in this regime is dominated by direct tunneling through the channel barrier since such tunneling is not very temperature dependent. Tunneling through the 22-nm effective channel length appears plausible for this device because the barriers are low, $< \sim$50 mV, and the effective mass of the lowest quantum level is quite low in the transport direction. Furthermore, the ideality factor of $\sim$3 for this FET increases the apparent tunneling inverse slope significantly above what it would be in a more ideal device.

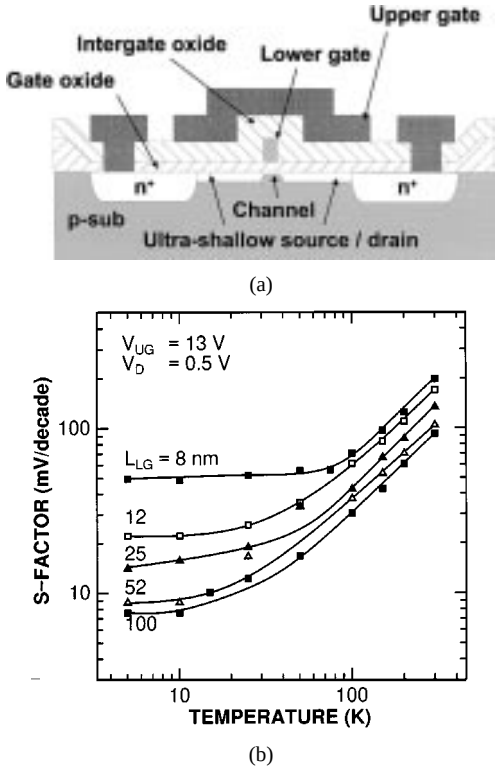


Fig. 13. (a) Cross-sectional view of 8-nm channel length EJ-MOSFET. (b) Temperature dependence of the inverse subthreshold slope at various channel lengths. Reprinted from [39] by permission of H. Kawaura.

On the theoretical device design front, recent work by Pikus *et al.* [42], [43] has shown that it should be possible to scale DG-FETs [see Fig. 3(b)] down to 8-nm channel lengths for logic and DRAM. These simulations use a ballistic transport model to predict device IV characteristics such as those shown in Fig. 14 for a Si channel 1.5-nm thick with 2.5-nm SiO_2 gate insulators on both sides of the channel. These particular curves do not include source-to-drain tunneling, but their later work does, showing that the effect becomes important at $\sim$8-nm channel length at 300 K. Using the three-layer generalization of the scale length theory [see (A5) in the Appendix], these FETs have $\Lambda_1 = 9$ nm so to allow for reasonable gate length and V_T tolerances in an FET without V_T rolloff compensation the minimum channel length should be $\sim$13 nm. However, if the design criteria for use in logic is only that there be sufficient gain, then these simulations show that such MOSFETs could be useful for logic down to about 8 nm or $L/\Lambda_1 = 0.9$. For DRAM, the most important thing is ON–OFF ratio and the simulations suggest that a sufficient ON–OFF ratio can also

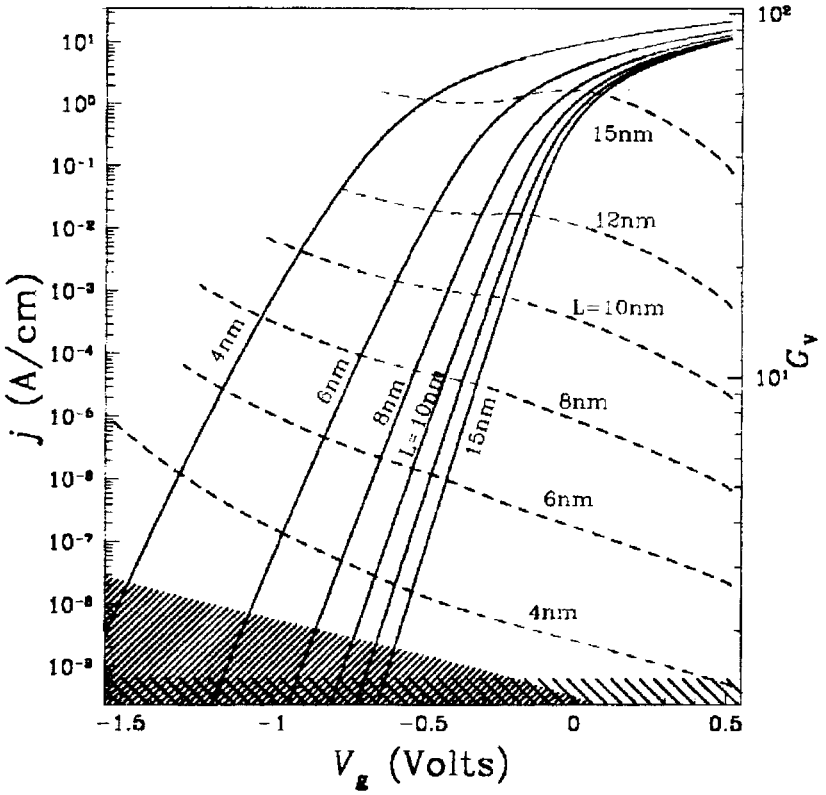


Fig. 14. Drain current and voltage gain $G_V = dV/dV_G$ (at constant current) versus gate voltage for very short channel DG-FETs. Reprinted from [42] by permission of K. Likharev and the American Institute of Physics, © 1997.

be obtained for channel length down to 8 nm, at which point source-to-drain tunneling becomes significant.

Finally, consider the ultimate limit for conventional bulk-like MOSFETs. According to Section II-B, the thinnest possible depletion depth for bulk FETs at maximum body doping is about 8 nm. Coupling this with a 1.2-nm very leaky gate insulator and 0.7-nm gate poly-Si depletion gives a scale length of ∼11 nm, which ought to make it possible to consider FETs with channel length of order 10 nm, in keeping with the preceding examples. Unfortunately, it will be very difficult to get the desired low-threshold voltages with this design because quantum confinement effects will raise V_T at least 200–250 mV at such high fields [6]. If a high-k gate insulator is not available, it will be even more difficult because the resulting high-ideality factor $\eta \simeq 1.5$ will cause a high V_T due to the built-in field and will necessitate a high V_T due to the low-subthreshold slope.

It appears possible to approach this regime, however, by cooling the CMOS chip to low temperatures and forward biasing the body. While low-temperature operation by itself does not help 2-D effects, some of the improvements in subthreshold slope can be traded off for a narrower gate depletion width to attain better control of short-channel effects. A forward body bias in this case is helpful in several ways. First, a forward body bias reduces the built-in potential and adjusts V_T to lower values, both directly and by decreasing the field and the quantum confinement energy. If the body is forward biased by 0.5 V, the depletion depths can probably be reduced to ∼5 nm, making it possible to achieve a 10–12-nm channel length. Second, a forward body bias lowers the reverse bias and therefore the field across the drain-to-body junction, hence suppressing the band-to-band tunneling current. Meanwhile, the leakage current of the source junction, although forward biased by the applied body voltage, is insignificant at low temperature as long as the body bias does not exceed ∼0.5 V.

The steeper subthreshold slope at low temperature allows V_T and, therefore, V_{DD} to scale further below their room-temperature limits given in Fig. 2(a). With the threshold voltage scaled to 0.1 V or so for 100-K operation, we estimate that it is possible to extend CMOS to ∼11-nm channel length with a 1.2-nm t_{ox} and 0.5 V V_{DD}. 2-D drift-diffusion simulations show inverse subthreshold slope of 40 mV/decade even for 1.5 nm t_{ox} at 100 K and 10-nm channel length, so achieving a low V_T with low OFF current appears quite feasible, but these simulations do not include source-to-drain tunneling current. Separate estimates indicate that this tunneling current will start to dominate the thermal OFF current somewhere in the 10–12-nm regime, thus creating the ∼11-nm channel length limit.

These design points by their definitions do not include any tolerances and, thus, serve as reference points for what may be possible if process variations could be completely controlled.

IV. Practical Limits on MOSFETs

As shown in the previous section, the ultimate theoretical limit on the size of a MOSFET is very small indeed. Unfortunately, the commercial use of FET technology is constrained by a variety of factors that (at least presently) preclude reaching the ultimate limit, except in one-of-a-kind devices. Perhaps the most important of these factors is the manufacturing reality of tolerances. These tolerances arise both from processing variations and from circuit conditions.

On the processing side, there are lithographic variations due both to exposure conditions and to photoresist variations. At the finest level, the molecules of the photoresist are discrete and may cause a certain level of fundamental coarseness. From a device point of view, the most important consequence of these lithographic variations is a random variation in the gate length. This variation occurs both from device to device within a chip, due to exposure nonuniformities, proximity effects, stochastic effects, etc., and as an average variation from wafer to wafer and chip to chip, due to imperfect control of processing conditions. All other aspects of manufacturing are subject to control tolerances, too, including layer growth or deposition thicknesses, etch depths and profiles, ion implantation conditions, and annealing conditions. None of the current manufacturing processes controls the exact atomic position of each dopant atom and this uncertainty by itself can lead to substantial $\mathrm{V_T}$ variations in very small FETs, as described Section IV-D.

There are also circuit related tolerances due to the capacitive coupling between signal lines. Since a computer has a very large set of possible states, there is a statistical distribution of noise coupled onto each signal line. For very low supply voltages, the logic state variations described in Section II-C are another source of such noise.

Both of the above forms of uncertainty must be accounted for in designing optimized devices and circuits. Several studies have been done at the circuit level characterizing the effects of these tolerance on $\mathrm{V_T}$ and $\mathrm{V_{DD}}$. The basic result is that optimized threshold and supply voltages must be raised somewhat to accommodate these effects compared to what they would be for perfectly controlled FETs with nominal characteristics [44]–[47]. Section IV-A and Section IV-B describe practical attempts to address the scaling limits of MOSFETs in the context of tolerances, Section IV-C addresses one of the sources of variations, namely discrete dopant effects, and Section IV-D briefly discusses power dissipation.

A. Bulk CMOS

Bulk CMOS has been the mainstream VLSI technology for the past two decades. Below 100-nm linewidth, however, CMOS design options are severely constrained by the fundamental issues of oxide tunneling and voltage nonscaling discussed in Section II. To explore in more detail the limit of bulk (and PD-SOI) CMOS scaling, we present a feasible design for 25-nm (channel length) bulk CMOS without complete scaling of oxide thickness and power supply voltage [27]. Such channel lengths can be achieved at a lithography generation of 75-nm resolution in year 2008 according to the ITRS roadmap [4]. Key issues such as gate work function, channel and source–drain doping requirements, poly-Si depletion effect, and nonequilibrium carrier transport in 25-nm CMOS are addressed. As discussed in Section VI, it may be possible to scale a little further than this, but only at extremely high power.

While straightforward 2-D scaling calls for a gate oxide around 1 nm for 25-nm MOSFETs [see Fig. 2(b)], direct tunneling leakage in oxide/nitrite gate insulators is very high for such thin insulators as already discussed, so we take $\mathrm{t_{oxCeq}} = 1.5$ nm, which is near the limit described in Section II-B. To maintain reasonable OFF currents on the order of 100 nA/μm for an integration level of 10^8–10^9 devices per chip, the room-temperature threshold voltage is kept at a minimum of about 0.2 V under the worst case conditions. The power supply voltage is set at 1.0 V, which represents a reasonable tradeoff among active power, device performance, and high field effects. With the nonscaled gate oxide and supply voltage, an optimized vertically and laterally nonuniform doping profile called the superhalo [16] is needed for controlling short-channel effects. Fig. 15 shows such a doping profile along with simulated potential contours for a 25-nm MOSFET. In principle, such a profile can be realized by ion implantation self-aligned to the gate edges with very restricted amount of diffusion. The highly nonuniform profile sets up a higher effective doping concentration toward shorter devices, which counteracts short-channel effects. This results in OFF currents insensitive to channel length variations and allows CMOS scaling to the shortest channel length possible. In the 25-nm CMOS design shown in Fig. 16, I_{off} is nearly independent of channel length variations between 20 and 30 nm. The superior short-channel effect obtained with the superhalo is shown in Fig. 17 compared with a nonhalo retrograde profile. Because of the nearly flat $\mathrm{V_T}$ dependence on channel length, superhalo allows a nominal device to operate at a lower threshold voltage, thereby gaining significant performance benefit: 30%–40% over nonhalo devices for 25-nm CMOS at 1.0 V [27]. It should be pointed out that DIBL, which is still present in superhalo devices, has only a minor effect on the delay performance for a given high-drain $\mathrm{V_T}$.

The above 25-nm device design does not require stringent scaling of junction depth. Fig. 17 shows that the $\mathrm{V_T}$ rolloff is rather insensitive to the vertical junction depth with only a slight change when the junction depth is doubled from 25 to 50 nm for the same halo profile. This allows the junction depth to decouple from the channel length, thus avoiding the high-resistance problem with very shallow extensions. The lateral source–drain gradient, however, is much more critical. As CMOS channel length is scaled down, the lateral doping profile of source and drain junctions should also sharpen in step and be kept abrupt on the scale of a fraction of the channel length. Otherwise, short-channel effects degrade rapidly [27]. This is because channel length is largely determined by the points of current injection from the surface layer (inversion or accumulation) into the bulk, which takes place at a source–drain doping concentration

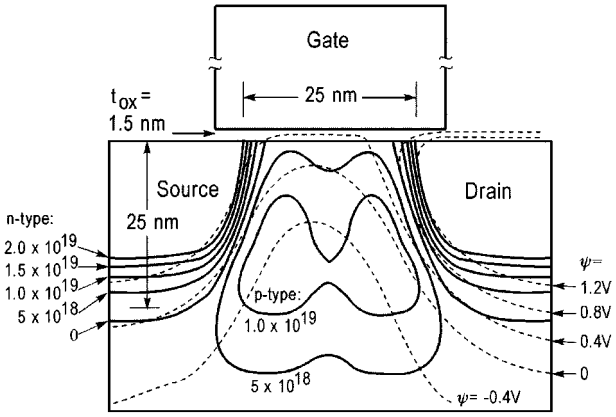


Fig. 15. Source, drain, and superhalo doping contours in a 25-nm nMOSFET design. The channel length is defined by the points where the source–drain doping concentration falls to 2×10^{19} cm^{-3}. Dashed lines show the potential contours for zero gate voltage and a drain bias of 1.0 V. $\psi = 0$ refers to the midgap energy level of the substrate. From [27].

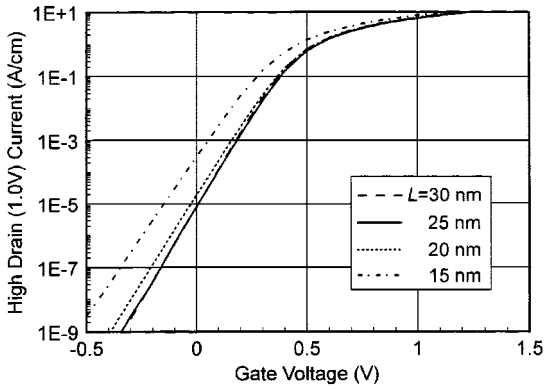


Fig. 16. Subthreshold currents for channel lengths from 30 to 15 nm. $I_{\rm off} = 10^{-5}$ A/cm (1 nA/μm) for 20, 25, and 30 nm devices. From [27].

of about 2×10^{19} cm^{-3} [48]. Any source–drain doping that extends beyond this point into the channel tends to compensate or counterdope the channel region and aggravate the short-channel effect. The abruptness requirements of both the source–drain and the halo doping profiles dictate absolutely minimum thermal cycles after the implants. Note that a raised source–drain structure may help making contacts, but does not by itself satisfy the abruptness requirement discussed here.

As discussed in Section II-B, a key issue with the high p-type doping level and narrow depletion regions in this 25-nm design is the band-to-band tunneling through the high-field region between the p-halo and the drain. For the peak field intensity (1.75 MV/cm) at high drain and zero gate biases shown in Fig. 15, the tunneling current density is on the order of 1 A/cm^2 (Fig. 9). This should not constitute a major component of the device leakage current given the narrow width of the high-field region, ~15 nm according to Fig. 15.

The threshold design in Fig. 17 assumes dual n^+/p^+ Si work function gates for nMOS/pMOS, respectively. A midgap work function metal gate would clearly result in

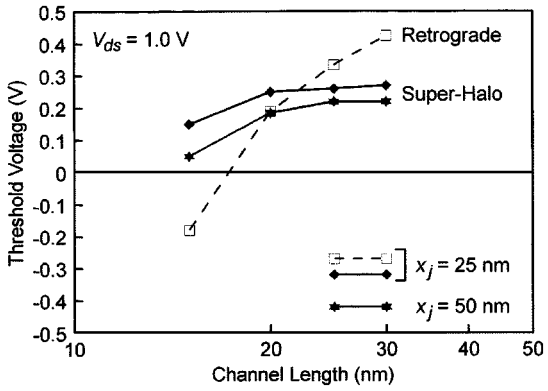


Fig. 17. Short-channel threshold rolloff for superhalo and retrograde (nonhalo) doping profiles. Threshold voltage is defined as the gate voltage where $I_{ds} = 1$ μA/μm. From [27].

threshold voltage magnitudes far too high for both devices [48]. With doped poly-Si gates, a frequently raised issue is the effect of poly-Si depletion on CMOS performance. Depletion effects occur in polysilicon in the form of a thin-space charge layer near the gate oxide interface, which acts to reduce the gate capacitance and inversion charge density for a given gate drive. The percentage of gate capacitance attenuation becomes more significant as the oxide thickness is scaled down. Actually, the net performance loss due to poly-Si depletion effects is much less severe than is suggested by C–V measurements. As it happens, the delay of intrinsic, unloaded circuits is only slightly degraded (~5%) because although poly-Si depletion causes a loss in the drive current, it also decreases the charge needed for the next stage. These two effects tend to cancel each other. For the heavily loaded case in which the devices drive a large fixed capacitance, the delay degradation approaches those of the ON currents (~15%). This can be compensated to some extent by using wider devices. On the average, the performance loss due to poly-Si depletion effect is about 10% for partially loaded 25-nm CMOS circuits with a 1.5-nm-thick oxide [27].

Extensive 3-D statistical simulations have been carried out on the effects of dopant fluctuations on threshold voltage for the above 25-nm device design [49]. Some of the details are presented in Section IV-C.

To evaluate the potential ON-state performance of 25-nm CMOS, detailed Monte Carlo simulations were performed using the simulator DAMOCLES [50]. Both n- and p-channel MOSFETs have been simulated, yielding low-output conductance high-performance I–V characteristics for both device types [27]. The transconductance exceeds 1500 mS/mm for this nFET, with an estimated $f_{\rm T}$ higher than 250 GHz. Transient Monte Carlo simulations were also done for a three-stage chain of 25-nm CMOS inverters. Fig. 18 shows the output waveforms. The estimated delay time is 4–4.5 ps, about three to four times faster than 100-nm CMOS operated at 1.5 V.

One way to go beyond 25-nm bulk CMOS is to cool the CMOS chip to low temperatures as discussed in connection to the 11-nm bulk MOSFET described in Section III. This is

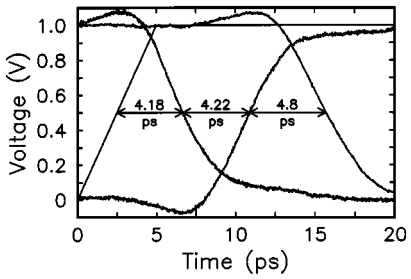


Fig. 18. Monte Carlo simulation of 25 nm CMOS inverter delay. pFET is twice the width of nFET. There is a third stage loading the output of the second stage.

feasible at least for high-end systems and offers the advantages of higher carrier mobilities and a steeper subthreshold slope that allows V_T and, therefore, V_{DD} to scale further below their room-temperature limits given in Fig. 2(a). To gain the most performance out of low-temperature CMOS, therefore, the threshold voltage should be tuned to lower values while maintaining the same OFF current as the temperature decreases [16]. If the 11-nm MOSFET can be realized as a worst case design point, then it should be possible to establish a nominal design point for low-temperature ($\sim$100 K) bulk CMOS at around 15-nm channel length. Of particular interest is the case when the p-type substrate (body to nFET) is biased at $V_{DD} = 0.5$ V and the n-well (body to pFET) is biased at ground potential. No extra power supply or on-chip voltage generator is needed.

B. Limits for DG-FET

The merits of the DG-FET have been analyzed by many researchers [11], [12], [29], [51]. There is a consensus that the electrostatic design of such FETs, with the gate completely surrounding the channel, is quite ideal and offers the potential to scale somewhat further than bulk devices [9]. Fig. 19 shows that DG-FET can be scaled to a channel length of about 20 nm [12] using 5-nm-thick Si and 1.5-nm gate oxide. This assessment is critically dependent on the assumptions that: 1) the silicon channel thickness can be controlled to within a reasonable tolerance; 2) the transport properties of DG-FETs are similar to those of single-gated bulk FETs despite the thin channel; and 3) the fabrication of DG-FETs does not impose additional constraints as compared to bulk FETs. This section addresses the three aforementioned assumptions.

Silicon channel thickness variations lead to threshold voltage variations from several sources: 1) short-channel effects due to the electrostatics of the device geometry; 2) quantization induced threshold voltage dependence on the silicon channel thickness; 3) threshold voltage dependence on the channel doping; and 4) random fluctuation of dopant number and dopant placement in a doped channel. Fig. 20 illustrates the threshold voltage variation due to short-channel effects, comparing the threshold voltage rolloff curves for channel thickness $\pm$15% from the nominal thickness. Analytically, by differentiating the Suzuki scale length [19], [52], the effect of channel thickness variation can be

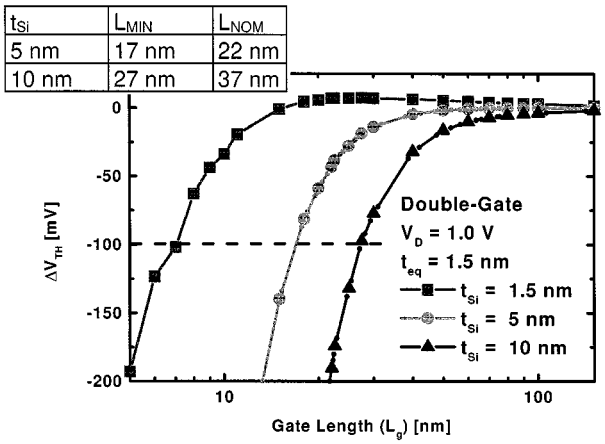


Fig. 19. Threshold voltage rolloff for DG-FET with equivalent gate oxide thickness (t_{eq}) of 1.5 nm and silicon channel thickness t_{Si} of 1.5, 5, and 10 nm. 1.5-nm thickness is not practical and is included here only as a reference. DIBL is taken into consideration by plotting the threshold voltage rolloff at a drain voltage V_D equal to the power supply voltage.

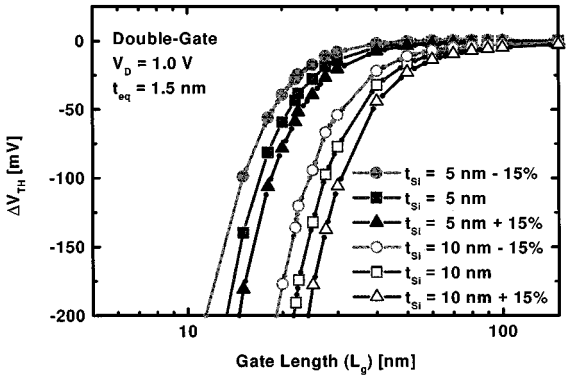


Fig. 20. Threshold voltage roll-off for DG-FET with equivalent gate oxide thickness (t_{eq}) of 1.5 nm. Silicon channel thickness (t_{Si}) is 5 nm and 10 nm with $\pm$15% variation.

converted to an equivalent variation in the channel length for a constant L/Λ_1 ratio [29]

$$\frac{\Delta L}{L} = \frac{2\epsilon_{Si}t_{ox} + \epsilon_{ox}t_{Si}}{4\epsilon_{Si}t_{ox} + \epsilon_{ox}t_{Si}} \cdot \frac{\Delta t_{Si}}{t_{Si}}. \tag{2}$$

In this way, the short-channel consequences of thickness variation can be accounted for as a part of the overall gate length variation budget. For high-k dielectrics, a more general formula can be found in the Appendix.

The quantization induced threshold voltage variation can be estimated from the particle-in-a-box approximation for the lowest subband [11], [12], [53], giving $\Delta V_T = -(h^2/4qm^* t_{Si}^2)(\Delta t_{Si}/t_{Si})$. Since this uncertainty grows rapidly with decreasing t_{Si}, it appears impractical to use a channel thickness very much below 5 nm [11], [12]. Consequently, although the $t_{Si} = 1.5$ nm case in Fig. 19 appears to offer very promising channel lengths down to $\sim$7 nm, it cannot currently be considered practical because it would have unacceptably large threshold variations. In the case where channel doping is employed to adjust the threshold voltage, the variation of

threshold voltage (for n-channel FET) with silicon channel thickness is $\Delta V_T = qN_A t_{Si} t_{ox}/2\epsilon_{ox}$ for accepters and $\Delta V_T = (qN_D t_{Si}/2)((t_{ox}/\epsilon_{ox}) + (t_{Si}/4\epsilon_{Si}))$ for donors. For thin channels, thin insulators and reasonable doping, it will be difficult to adjust the V_T by more than $\pm$100 mV. Random dopant fluctuation accounts for about 20–50 mV (one sigma) [49], [54] for practical doping levels employed to set the threshold voltages. Taking the above four factors into account, the tolerance for a $\geq$5-nm silicon channel thickness needs to be about 10%.

This silicon channel thickness tolerance requirement is quite stringent (about 0.5 nm for a 5-nm channel) considering the present state of the art for thickness control in SOI materials. Thickness tolerance of SOI wafers are typically $\pm$2–5 nm over an 8-in wafer. Bonded wafers typically have better tolerances than SIMOX wafers, both in terms of global thickness uniformity and local thickness variations (roughness). Over a smaller area (5-in diameter), bonded wafers show variations of $\pm$1 nm. This suggests that it may be possible to control the thickness to the required tolerances as technology for making bonded wafers progresses.

There have been few reports and predictions on the transport properties of short-channel DG-FETs [11], mostly due to the difficulty of modeling the physics precisely including the full 2-D quantum solution in the channel [55], [56] and the effects of dynamic switching [57]. There have been theoretical speculations on degradation of phonon-limited electron mobility in ultrathin silicon channels [58]–[60] due to the electronic structures of the confined thin silicon channel. Experimental verification of such degradation has not been made. Little has been said about carrier mobility at high normal fields where the FET operates.

Generally, the experimental data show mobility decreasing rapidly below 10 nm (see Fig. 21). However, most experiments on mobility for thin silicon channels reported in the literature contain too many uncertainties to be conclusive. Toriumi *et al.* [61] attributed the mobility reduction to Coulomb scattering from the interface traps at the back interface of the thin SOI. Choi *et al.* [62] ascribed the mobility reduction to silicon film stress for the thin silicon channel. Ernst *et al.* [63] showed mobility reduction, but did not offer any possible explanations. In all these experiments, the silicon channel was thinned by oxidation from a SIMOX wafer. This procedure introduces uncertainty due to the quality of the back interface. The source–drain series resistance tended to be high, which introduced more uncertainty to the measurements.

The fabrication of the ideal DG-FET is extremely difficult [54]. The "ideal" DG-FET should have [64]: 1) a uniform silicon channel, thin compared to the channel length, with $t_{Si} \leq 0.4L$ (see Section VI); 2) a thick source–drain fan-out structure to reduce the series resistance; and 3) top and bottom gates that are perfectly aligned to each other and to the source–drain dopings and fan-out in order to reduce overlap capacitance and series resistance of the ungated region. It may also require metal gates with specifically chosen workfunctions in order to obtain the desired threshold voltages since doping can only shift the V_T by $\sim$100 mV and it would be preferable not to require any doping at all to avoid discrete dopant effects. Conventional layer-by-layer-type fabrication techniques, which have served the microelectronics industry well for the past 25 years, are difficult to apply to the DG-FET structure that is somewhat 3-D. Various methods have been attempted, including selective epitaxial growth through a tunnel [64], forming a vertical silicon channel with side gates [65], [66], and wafer bonding with the channel and gates in place followed by selective epitaxial growth of the source–drain fan-out [67]. While these experiments generally show high series resistance and lower ON-current than expected, further innovations and perfection in the fabrication techniques should improve device characteristics in the future.

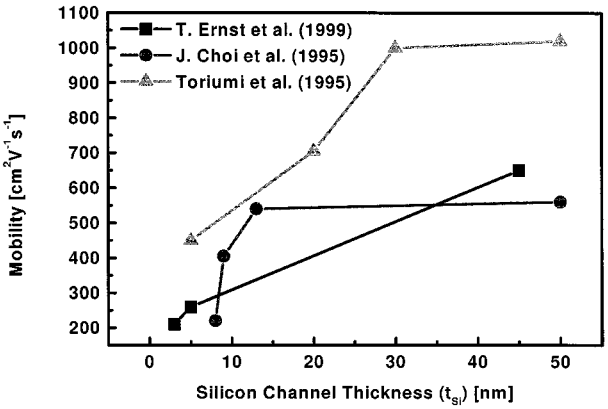


Fig. 21. Experimentally measured electron mobilities in thin silicon channels drop substantially below about 10-nm channel thickness. Lines are visual guides and do not suggest trends in the data. Mobility data of Choi *et al.* [62] and Toriumi *et al.* [61] are the peak mobility at low effective fields. Electric field corresponding to the mobility data of the Ernst *et al.* [63] was not specified in their paper and is presumed to be the low field mobility.

C. Doping Fluctuations

As was already mentioned, one of the potentially significant sources of variation in MOSFETs at the limits of scaling is randomness in the exact location of dopant atoms. Although the average concentration of doping is quite well controlled by ion implantation and annealing processes, these processes lead to randomness at the atomic scale in the form of spatial fluctuations in the local doping concentration, which in turn cause device-to-device variation in MOSFET threshold voltages. These fluctuations were anticipated long ago [68], but at the time most FETs had sufficiently many dopants that it was not a genuine problem. Since then, however, the number of dopants in the depletion region of an FET has been decreasing steadily with scaling, as illustrated in Fig. 22. The decrease has been roughly in proportion to $L_G^{1.5}$ due to the incomplete scaling of the electric fields, so that we are now into a regime in which the smallest FETs have fewer than 1000 dopants determining the threshold voltage. Since fluctuations in dopant number have a standard deviation equal to the square root of the number of dopants, in keeping with Poisson statistics, the $\pm 3\sigma$ bounds shown in Fig. 22 become extremely large by the time channel lengths reach 25 nm.

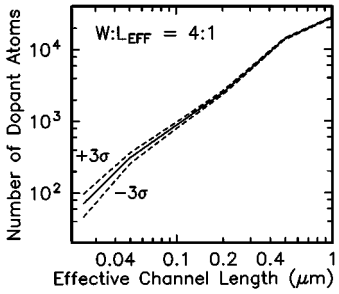


Fig. 22. Number of doping atoms in the depletion layer of a MOSFET versus channel length [49].

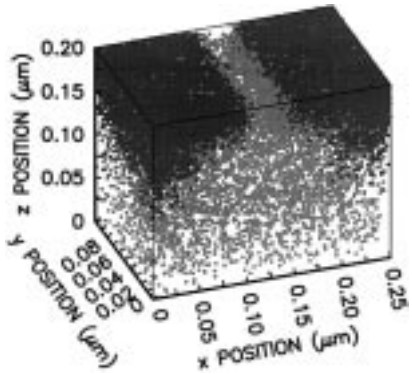


Fig. 23. 3-D perspective plot of the dopant atoms in a 25-nm MOSFET. Darker dots are donors and lighter dots are accepters. From [79].

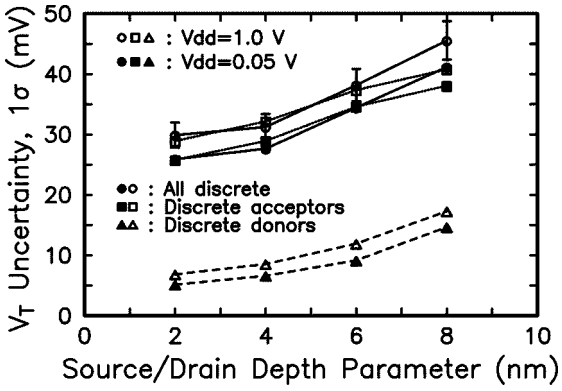


Fig. 24. Threshold uncertainty due to dopant fluctuations versus depth parameter, d, of the source–drain extension implants, showing the separate contributions of the donors and the accepters. The source–drain doping profile is a Gaussian, peaked at 10^{20} cm^{-3} at the surface, with standard deviation d in the vertical direction and $0.7d$ laterally. The junction depth $x_j \simeq 2.2d$. All widths are 50 nm, with 100 realizations for each point. From [49].

Many workers have investigated the effects of these doping fluctuations on the V_T of MOSFETs. The first model, proposed by Keyes [68], was an analytic approximation involving a percolating path from source to drain and has served as a basis for more recent analytic models, e.g., [69] and [70]. Various workers have also used 2-D numerical simulations [70]–[72], but the most quantitatively accurate work uses stochastically placed dopants in full 3-D MOSFET simulations to fully resolve the effects of dopant placement [49], [73]–[75]. Fig. 23 shows an example of such a doping configuration for the 25-nm MOSFET design described in Section IV-A. This particular example was created by a program that can analyze doping fluctuation effects for arbitrary doping profiles by associating a random number with every Si atom site in the entire device simulation volume. For each atom, the random number is compared against the local probability of a dopant atom (determined from the continuum doping concentration) to decide whether that atom is a dopant. These dopants are then snapped back to the simulation grid [49].

As an example of the results of such 3-D simulations, Fig. 24 shows the dependence of V_T uncertainty on source–drain depth for the 25-nm bulk MOSFET design. The threshold uncertainty increases with increasing junction depth because of the increasing body doping needed to maintain a fixed OFF current of 1 nA/μm. By separately simulating the effects of discrete donors or discrete accepters, the simulations also show that the effect of discreteness in the source–drain is usually negligible. Stochastic simulations also confirm the analytically predicted result that highly retrograde channel doping profiles can yield significantly ($>2\times$) lower σ_{VT}s than uniformly doped channels [49], [75]. This is because the doping fluctuations are moved further away from the channel and closer to the body and so have less effect since they are screened by the free carriers in the body [76].

Most importantly, these simulations reveal the magnitude of σ_{VT} for MOSFETs at the limits of scaling. The 25-nm designs, in particular, have $\sigma_{VT} \sim 7 - 10/\sqrt{w}$ mVμm$^{1/2}$, where w is the width. Even for idealized retrograde doping this value does not fall below $\sim 5/\sqrt{w}$ mVμm$^{1/2}$. In addition, Asenov *et al.* [77] have shown that quantum confinement effects add another $\sim$24% to these uncertainties. This means that high-performance logic devices which tend to be wide may only have a few extra millivolts of V_T variation, which would be lost amidst the process-induced variations, but small FETs such as those in SRAM cells may have σ_{VT} as high as 40 mV, which is sure to be a problem for large SRAMs [78] in which variations up to $6\sigma_{VT}$ or higher can be expected.

It is not yet clear how this SRAM yield problem will be met, but it is clear that MOSFET design will need to take into account dopant fluctuations by choosing doping profiles that reduce the problem. If channel doping profiles can be adequately engineered, published projections [9] show that it should be possible to meet the SIA roadmap requirements for σ_{VT} out to at least the year 2012, but it is not clear that these requirements are sufficient to guarantee circuit functionality at the intended supply levels. As in other cases, the DG-FET may have an advantage: since (under some conditions) it does not require as much doping to obtain the desired threshold, its fluctuations may also be lower [79].

D. Power Density

Power density is an important application issue, but not a fundamental limit. It was demonstrated in 1981 that nearly 1 KW/cm^2 could be removed from a Si wafer [80] by forcing liquid coolant through channels etched into the back of a Si

Table 3
Application Classes and their Minimum Equivalent Oxide Tunneling Thicknesses t_{oxTeq}

Application Class	P_{tot} (W/cm^2)	P_{ox} (W/cm^2)	V_{DD} (V)	J_{ox} (A/cm^2)	t_{oxTeq} (nm)	T (^{o}C)	$F_{ox,10yr}$
High Performance	100	10	1	667	1.2	85	0.25
						-40	$3x10^{-3}$
Desktop	10	1	1	67	1.4	85	10^{-2}
Short battery life portable	$50x10^{-3}$	$5x10^{-3}$	1	0.33	1.8	40	$3x10^{-7}$
Long battery life portable	$50x10^{-6}$	$5x10^{-6}$	1	$3.3x10^{-4}$	2.4	40	10^{-13}

P_{tot} is the total active power, P_{ox} is the power due to gate insulator tunneling, J_{ox}^{max} is the tunneling current density through the gate insulator, T is the operating temperature, and $F_{ox,10yr}$ is the estimated probability of an oxide failure in ten years of operation. Assume that 10% of P_{tot} can be allocated to gate leakage. 3% of the chip area is oxide, of which half is leaking at any given time. t_{oxTeq} is derived from Fig. 7, using J_{ox}^{max}. It is assumed that standby-mode dissipation can be eliminated by powering down unused circuitry and that dynamic circuitry may well be precluded by such high leakage. Since the low power cases must have higher V_Ts to reduce OFF current, the supply voltage has been kept at 1 V to maintain performance, although a somewhat better optimum is probably possible.

wafer and even more would probably be possible with further engineering effort. It is true that high switching activity circuits at a density commensurate with the source–drain tunneling-limited FET size can probably reach a power density $\sim$1 KW/cm^2 for small macros, but by judicious choice of circuitry and system architecture, such dissipation can usually be averaged down by other macros that are not so actively used. Consequently, practical limitations on power density are much more important than fundamental limits because there are many applications for which such a high dissipation is unacceptable, necessitating much more constraining scaling limits, as discussed in Sections V-A and VI.

For SOI device designs, the 100$\times$ worse thermal conductivity of SiO_2 creates additional concerns. At high-power density, SOI devices can experience unacceptable local temperature rises, especially under dc measurement conditions. For realistic buried oxide thickness and switching device duty factors, however, it can be shown [9] that these temperature rises are quite contained and should not impose a fundamental limitation.

V. Application Dependencies

A. General Considerations

Having looked at some of the fundamental, theoretical, and practical limits to scaling, we now consider in more detail the application dependencies of some of these limits. One of the most important "limit"-related issues is the power dissipation associated with leakage currents. Active power is generally determined by $CV_{DD}^2 f$, where f is the clock frequency, and can be adjusted by changing V_{DD} and f, but leakage currents depend to a large extent on device design.

It is widely understood that V_T should be set differently for different applications to control the subthreshold leakage dissipation. As scaling proceeds further, however, it is important to understand that power dissipation associated with gate tunneling current also needs to be managed. Depending on the application, different amounts of tunneling dissipation can be tolerated and this translates into different minimum t_{oxTeq} for different applications. Table 3 illustrates the results of such an analysis [81] using the aggressive premise that a full 10% of the active mode power can be dissipated as gate leakage. The use of 10% is an engineering tradeoff estimate based on the idea that power usage should be reasonably balanced (in this case, 2/3 active switching power and 1/3 passive steady–state dissipation, assuming 20% goes to subthreshold current). High-performance dynamic circuits might be impacted by this leakage level, but static CMOS should still function. The table illustrates very clearly that, contrary to the ITRS99 assumptions [4], the thinnest insulators are not suitable for low-power applications since they leak too much.

One immediate consequence of this analysis is that there cannot be a single "end of scaling," but rather, there will be a range of different device designs for different applications, utilizing a range of gate insulator thicknesses, whatever that best insulator turns out to be. Note also that if a better insulator than SiO_2 cannot be found, we are already at the end of gate insulator scaling for some applications. The 10-yr reliability estimates in the last column are for pure SiO_2 gate insulators only since there is very little data on other materials. They are intended to provide a rough order-of-magnitude illustration of the reliability situation for ultrathin oxide [82]. They assume a total thin oxide area of 20 mm^2 for each case stressed at a 50% duty factor, i.e., for a cumulative total of five years. The potentially high failure rate of 1.2-nm oxide at elevated temperature suggests that low-temperature operation of high-performance processors may be important not only for speed, but also to achieve acceptable reliability.

The following sections explore in more detail scaling limit issues for some of the more important classes of applications.

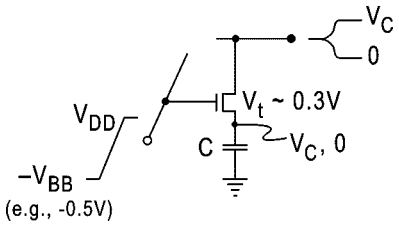
B. DRAM

1) Scaling Challenges for DRAM: Although DRAMs and microprocessors share a common technology base, their product requirements and technical challenges are considerably different. DRAMs are driven by the goal of reducing cost/bit in each generation, which has been achieved by having greater density and more efficient production of larger chips with larger starting Si substrates. DRAMs generally use higher internal voltage to store maximum charge in the memory cells and there is a requirement for low leakage to minimize the refresh activity. Microprocessors are driven by the demand for higher and higher speed. The highest performance processors are now going to lower voltage levels to keep the much higher chip power within reason. Leakage current does not need to be as low as for DRAMs, except in some battery-powered applications.

Because of these requirements, the CMOS technology used in microprocessors has been scaling to small dimensions at a faster pace than that used in DRAMs. As scaling of both DRAM and microprocessors continues into the 21st century, clearly there will be a drive to integrate them at the chip level to increase the memory bandwidth as demanded by the faster processors. This will place a premium on DRAM speed for some applications while density and cost will remain the driver for the large-volume memory applications.

The steady progress in DRAM up to the present has been driven by dimensional scaling of the devices and wiring plus continuous improvements in the basic memory cell to achieve a more compact areal layout, such as the use of trench or stacked capacitors. While there is some opportunity for further scaling of DRAM devices, it is becoming very difficult to find still more compact cell arrangements. Here we discuss first the scaling of devices and voltages in DRAM and then the challenges and possibilities for improved cell concepts and structures.

2) DRAM Devices and Voltage Scaling: DRAM memory chips, for a given lithography capability, now use longer channel lengths and higher voltage levels on the gate compared to the performance-oriented logic devices in order to store more charge on the capacitor. Most present circuits achieve a voltage difference V_C on the capacitor which is about 1.5–1.8 V less than the peak voltage applied to the gate of the memory cell devices. This is because of the need for a high-threshold voltage V_T in the memory cell devices ($\sim$0.8 V) to prevent subthreshold conduction of charge from the capacitor to the bit line at times when the bit line is at a low voltage and because of body effect, threshold tolerances, and signal required to turn on this device adequately to write the high-level V_C into the capacitor. In the future, as DRAMs are scaled to smaller dimensions, the voltage that can be applied to the memory devices will follow a path similar to logic devices (but delayed in time) because those devices are at maximum field strength for gate-oxide reliability in any given generation [83]. Therefore, the stored voltage on the capacitor will shrink rapidly as the voltages are scaled down unless a better technique is found. Also it is very difficult to achieve such a high V_T in these scaled devices, at least with the usual poly-Si gates.



(a)

Device Voltage, V_{DD}	Capacitor Voltage, V_C
2.5	1.5
1.8	1.0
1.5	0.8
1.2	0.6

(b)

Fig. 25. Negative wordline voltage technique for DRAM memory cells. (a) Memory cell schematic. (b) Scaling path for V_C versus V_{DD}.

A basically better arrangement, which has been used in some DRAMs, is shown in Fig. 25 [84]. It uses a lower threshold device that is easier to make and allows a higher voltage level V_C to be stored on the capacitor for a given wordline up level V_{DD}. Leakage of charge back through the device is shut off by keeping the turned-off gate negative compared to the lowest source voltage on the bit line or on the capacitor, which is zero in this case. When the transistor is turned on to write V_C or zero into the capacitor, the gate is stressed to V_{DD} for the case of zero on the bit line. The table in Fig. 25 lists the estimated capacitor voltage that can be stored as a function of the maximum device voltage V_{DD}. This assumes that the body effect, threshold tolerances, and signal required on the device all scale down as the device is scaled to operate at the lower V_{DD}.

Up to now, DRAMs have commonly used a thinner equivalent t_{ox} for the storage capacitors in the memory cells compared to the t_{ox} used for the gate insulators. This has tended to maximize the charge stored on the capacitors considering the lower voltage stress on them. Sustaining this trend with further scaling appears to be challenging. Rather than simply making the capacitor insulator thinner, which will soon lead to high tunneling current, this suggests that higher dielectric constant materials will be needed. The leakage current requirement for the capacitor is quite stringent because of the large area involved. If SiO_2 were used, the limiting thickness would be about 3 nm for trench capacitor structures ($\sim 10^{-6}$ A/cm^2 leakage). The commonly used nitride–oxide composite can be scaled to a somewhat thinner equivalent oxide thickness than that, perhaps 2.5 nm.

The leakage current requirement for DRAM is also a significant challenge for scaling of the cell transistor. Although DRAM devices are properly turned off with the scheme of

Fig. 25 there is a concern that drain junction leakage at the gate edge (also known as gate-induced drain leakage, or GIDL) may discharge the capacitor [85]. This effect is greatly alleviated when the junction voltage approaches the Si bandgap. However, tunneling current in the drain-body junction due to heavy body doping is a concern for devices using a heavy "halo" or pocket implant. The tunneling current density that can be allowed in the junction is about 10^{-4} A/cm^2. This is reached for a background doping density of about 3×10^{18}/cm^3 for an ideal junction [26], but may require a much lower concentration because of defects [34].

Tunneling current through the gate insulator is also a concern. With the gate of the array transistor biased to a negative value, relatively few electrons can tunnel from the gate into the weakly inverted channel and only a portion of these will flow to the drain. The potentials are favorable to tunneling in the gate–drain overlap region, but the gate–drain insulator thickness can be increased relative to the t_{ox} in the channel region by a "bird's beak" created in the gate-reoxidation process. It appears, therefore, that the most critical region is the boundary between the "bird's beak" and the channel, where the oxide is thinner but the channel potential is still near that of the drain. In this region a current density of 10^{-4} A/cm^2 can be tolerated, which corresponds to a t_{ox} around 2.5 nm for the biases of interest.

3) Future Directions: From the above considerations it appears that DRAMs can be scaled to effective channel lengths of 50–100 nm if a direction for increasing the capacitance/unit area is found. However, it will be very difficult to continue the past trends in compacting the memory cell, i.e., decreasing the size normalized in lithographic squares. The present approach, which gives about eight squares per cell, can only be improved significantly by going to a vertical structure for the device as well as the capacitor. Such structures have been proposed for some time, but no ideal solution has been found. The approach of Gruening *et al.* [86] looks interesting and appears to be capable of reducing the cell area to six squares.

Undoubtedly, progress in scaling DRAM will continue for some time to the degree that lithography advances and the increased density on a chip will allow new systems approaches. It seems likely that today's conventional DRAM will be embedded with microprocessors on the same chips and speeded up by utilizing the microprocessor devices and architectural improvements [87], while chips that perform only the memory function will continue to emphasize density, slower speed, and perhaps nonvolatility. Thus, DRAM may evolve in more than one direction in the future.

C. SRAM Limits

For most large logic applications, the SRAM is a highly critical component. The SRAM cache access time is a critical part of the system access time. A large fraction of the transistors in a modern processor chip are in SRAMs, impacting the density and standby power of the entire chip. There are several different types of SRAM in use in these systems, so it is important to understand their different requirements.

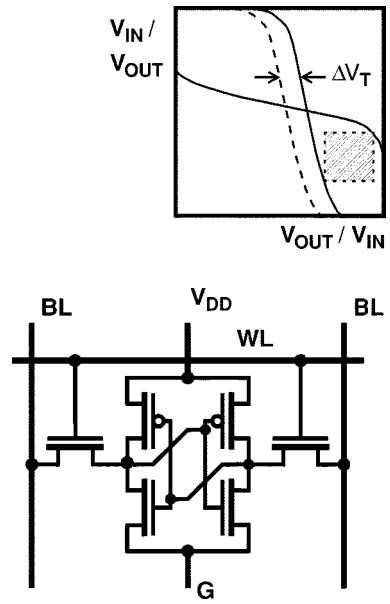


Fig. 26. Six transistor Static RAM cell. Cell is addressed by the wordline (WL) and the data is read out via the bit line pair (BL). Inset shows a set of superimposed transfer curves for the two halves of the SRAM cell showing how the noise immunity (shaded box) is affected by a threshold voltage shift in one of the halves.

A typical high-speed SRAM uses the standard six-device cell configuration as shown in Fig. 26. The typical area of these high-performance SRAM cells ranges from about 100 to 200 lithographic squares. This is to be compared with only eight squares for a DRAM cell and even smaller areas for some nonvolatile memory cells. Why, then, does the industry tolerate such an inefficient use of silicon area? The main factors determining this are high speed, compatibility with logic, and the use of caching. These factors mean that performance has much more weight than density, at least in the first level (L1) cache designs since demands for higher density can be shunted up the memory hierarchy. To appreciate the memory hierarchy in the context of technology evolution, consider Fig. 27. This figure shows how a larger part of the memory hierarchy is placed on-chip as the technology progresses. Stand-alone SRAM, which today has important uses both for cache and in signal processing systems, will probably become part of a larger on-chip system in the future. For instance, a recent microprocessor unit (MPU) [88] features 32-kB instruction and data caches as well as an on-chip 256-kB L2 cache. In the future the L1 cache is likely to increase fairly slowly in size to maintain its high speed, whereas several megabytes of L2 cache might eventually be used and perhaps even additional levels of hierarchy would be incorporated on chip.

Practically, this means that the speed-critical SRAM, which must use the general logic technology, becomes a smaller part of the total chip and issues of power dissipation become less important for this part. These SRAMs become merely an extension of the logic technology with no special consideration as to density, standby power, etc. Fairly large subthreshold currents can be tolerated; for instance, a 64-kB cache ($\sim 10^6$ current paths) at a power supply voltage of 1

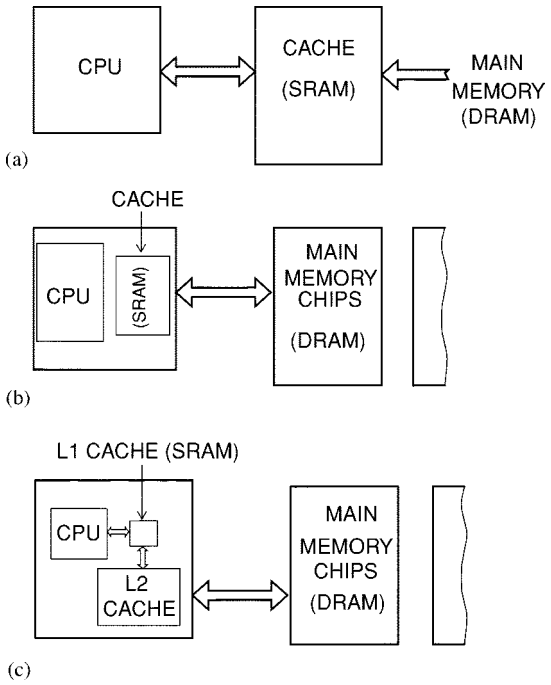


Fig. 27. Evolution of the memory hierarchy with increasing integration level (a) CPU and Cache on separate chips. (b) CPU and Cache on the same chip. (c) CPU + 2 level cache on one chip.

V can tolerate $\sim$1 μA/μm of leakage current for 1-W power dissipation, which is consistent with low V_T FETs.

The picture is very different for the L2 cache, which now has a relaxed speed (latency) requirement compared to the L1 cache, but is much larger. Many tradeoffs can be made for speed versus density and completely different technologies can, in principle, be incorporated such as DRAM, floating gate and thyristor memories [89], which use silicon technology but add processing steps to the CMOS logic process, or memories such as ferroelectric and magnetic RAM, which introduce new materials, with the cost being the determining factor. Given the design latitude with emphasis on density rather than speed, other SRAM approaches than the 6-T cell such as 4-T and 5-T versions or cells using resistor loads placed above the cells (or preferably thin-film transistor loads) to reduce standby power might find their place. At the cost of the added technology involved in stacked load devices, SRAM cell size can be reduced to the 50–100 square range, but this is currently practiced only for stand-alone SRAM chips. The 4-T cell has reduced static noise margin (on read) as compared to the 6-T cell, especially at low voltages [90], but there will still be a great incentive to use these designs given the increased density, where the noise margin issue can be countered with appropriate choice of voltage (see below) and a less demanding speed requirement.

A conventional SRAM L2 cache would still be the cheapest choice for many applications since it does not add any processing cost if it uses the general logic technology although even here, modest technology enhancements to improve density, such as buried contacts, may prove to be cost effective. Cell layouts may be used that maximize density rather than speed by using narrower transistors and using the multilevel wiring capability for the crossovers. Further density improvement can also be obtained by repartitioning the SRAM to reduce peripheral circuitry. The L2 SRAM would need high-threshold voltage transistors in the cells to minimize device leakage as well as a thicker gate oxide to reduce gate leakage. Such options are already being incorporated into today's technology and will continue to be necessary in future scaled logic technology. The L2 cache would most likely also use longer channel length FETs to reduce V_T rolloff variations. The longer channel length would not impact the cell size severely because all of the other groundrules will remain unchanged. The higher threshold cells will also need a higher power supply voltage than the general logic and a higher swing from the wordline drivers to adequately turn the transfer devices both on and off [91]. To save dynamic power, the bit lines and sense amplifiers may run at the somewhat lower voltages of the general logic [92]. Dual or multiple power supplies will probably be available from the future technology for these and other purposes.

The above scenario, which is very likely, leads to the possibility that most of the standby current leakage requirements of the chip could be shunted to a less demanding technology in terms of threshold voltage, channel length, gate leakage, and reliability, leaving more freedom for the logic technology to be optimized for performance.

Potential issues for future SRAMs are soft errors due to scaled-down voltage and capacitance and hard fails due to threshold voltage variability of the small in-cell FETs due to random dopant fluctuations and other processing issues. Burnett *et al.* [78] investigated cell stability under random dopant fluctuations and showed how this leads to increased voltage requirements for the cell in scaled technologies. In practice, redundancy techniques can and are being used to reduce the impact of the hard fails [88] and error-correction techniques can reduce the soft error rate to an acceptable level. Alternate technology choices such as thin SOI with or without a double gate could reduce soft errors through decreased collection volume and hard errors by eliminating (or reducing) the body doping.

D. Low-Power Applications

There is a steadily growing market for low-power applications of CMOS technology and it is the battery-powered nature of most of these applications that particularly creates the low-power constraint. To achieve good battery life, these circuits simply cannot dissipate very much power. Roughly speaking, these are circuits that consume less than 1 W/cm^2 with a subgroup of ultralow power circuits in the range below 1 mW/cm^2. Higher power applications are discussed in the next section.

Low-power constraints fall into two broad categories: those that relate to active mode power dissipation and those that relate to dissipation in the quiescent state. Some types of applications are primarily sensitive to active power considerations, since they are switched off when not in

use. Other applications may be turned on almost all the time, but rarely ever actually compute anything and so are more concerned with the quiescent power dissipation. Since MOSFET design limits are different for these two cases, they need to be considered separately.

Reducing active power begins at the top. The most effective place to reduce power dissipation is almost always at the highest level of the problem definition. Redefining the problem, the architecture, the algorithms, and/or the protocols can often save several orders of magnitude in power dissipation. The development by Meng *et al.* [93] of a portable video-on-demand chip set using only 10 mW is an example of this.

At the device design level, the important low-power variables are the threshold voltage, the gate leakage current, and the device size, which largely determines the body-to-drain tunneling dissipation. For current generations of technology, the latter effect is not usually significant, but at the limits of scaling, it should become quite important. For active mode dissipation, these parameters offer strong tradeoffs between speed (at low V_T, thin oxide, and small devices) and low power (at higher V_T, thicker oxide, and larger devices). This tradeoff occurs because all three variables tend to simultaneously increase the circuit's speed and its dissipation during the time it is not switching.

The optimization of the V_T and V_{DD} tradeoffs for low power has been well studied [44]–[46], including the effects of process and supply variations, which are quite important. The study by Frank *et al.* [46], for example, shows that even in the presence of realistic variations, the optimum supply voltages can readily drop below 1 V and can reach 0.5 V under some conditions (high switching activity and switching speed target 5–10× slower than the maximum technology capability). The optimum value for V_T increases for slow circuits to reduce static dissipation and increases by 20–100 mV when the tolerances are doubled from their nominal (realistic) values. The dependence of the optimum design points on activity factor and logic depth is illustrated in Fig. 28. These particular optimizations are for 0.1-μm static CMOS arithmetic circuits and each point in the figure represents an independent optimization of both the supply voltage and the threshold voltage. Optimization of the gate length at the 0.1-μm generation was a weak effect, but at the limits of scaling it will be very important because of body-to-drain tunneling, as discussed in Section VI. As shown, the optimum nominal threshold voltage depends strongly on activity factor and logic depth, so that in the limit of minimum power, a wide range of V_Ts are needed to satisfy the requirements of a range of applications.

The optimization of oxide thickness for low-power consumption has only recently become a concern and so has not received significant study. The rough estimate used in Section V-A of 10% of the total power allocated to gate leakage seems reasonable as a first pass. Based on this estimate, as indicated in Table 3, the likely minimum t_{oxTeq} varies from ~2.6 nm for ultralow-power applications to ~1.7 nm for moderate low power, although it could go somewhat thinner

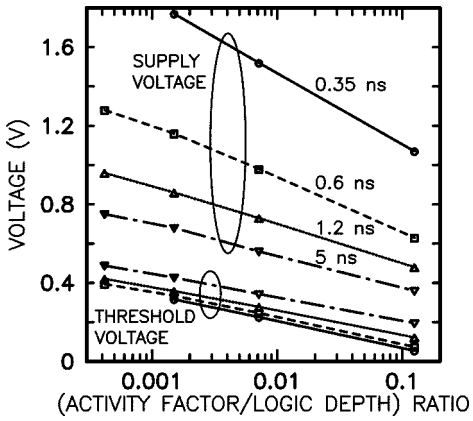


Fig. 28. Plot of supply voltage and threshold voltage versus the activity factor-to-logic depth ratio for four different delay constraints. Logic depth $= 4n_{stages} = 48$, 40, 28, and 8 for the data points from left to right across the plots. Threshold voltage here is the gate voltage at which $I_{on}^{1/2}$ linearly extrapolates to zero. Data is from [46].

for cases in which the optimum supply voltage is well below the 1 V used in the table.

Quiescent power constraints are often much lower than the optimized background dc dissipation during active mode. This means that the optimum active mode FET design cannot satisfy the quiescent constraints. This problem can be addressed in at least three different ways: 1) one can use higher V_T FETs throughout the design, so that the quiescent power requirements will be met; 2) one can dynamically change the body voltage (or backgate voltage for double-gated structures) to switch between a high-V_T quiescent state and a lower V_T active state; and 3) one can put series switches in the power supply to turn off inactive circuit blocks.

The drawback to using higher V_T FETs throughout is that one must then raise the supply voltage to maintain adequate performance in the active mode. One must also use thick enough oxide that its quiescent leakage satisfies the constraint, requiring still higher V_{DD} to retain performance. The device size must also increase to support the increased voltage without increasing the body-to-drain tunneling currents. Thus, the active mode is likely to consume much more than the minimum required power, making this an inefficient option unless extremely little active power is ever required.

Dynamically adjusting V_T using the body voltage is an interesting option which has also been suggested as a way of compensating out process-induced threshold variations. The biggest disadvantage of this approach is that it only eliminates dissipation due to subthreshold leakage, but does not take away the gate and body-to-drain leakage dissipation. There is also the question of whether body junction leakage and bias generation power may not exceed the subthreshold dissipation one is trying to suppress.

The third option appears to be the best for logic circuits, since it eliminates all three forms of leakage dissipation, leaving in their place only the leakage dissipation associated with the series switch. The disadvantage here is in

controlling and powering the series switch, but this appears feasible. Another consideration is that it is not permissible to power down the latches and SRAM if they are holding data that must be retained. Data-holding circuits seem to have little choice but to use high-threshold voltages in order to suppress quiescent dissipation. This line of reasoning strongly suggests that low-power chips will require at least two threshold voltages, one low and optimized for logic, and one high and optimized for holding data. It should, however, be possible to reduce the V_{DD} applied to data-holding circuits during quiescent mode, which would reduce their tunneling currents.

E. High-Performance Applications

The last dozen years have seen tremendous growth in a specialized, but significant, application of CMOS logic, namely that of high-performance CMOS. This application is defined by products in which raw switching speed is the primary goal of the process technology. Active power dissipation is also a consideration, albeit usually a secondary one. The role played by subthreshold leakage divides this application space into a few subcategories which we discuss explicitly. Furthermore, as scaling continues into the sub-100-nm regime, we expect that gate oxide leakages will also place limits that differ by application in a similar manner. The high-performance CMOS scaling trends since 1990 were shown in Fig. 2 from which it is clear that as silicon manufacturers near production of the 0.13-μm node of the ITRS roadmap [4], the industry is on the verge of a regime where there is no room to continue the past trends for threshold voltages.

1) Worst Case V_T Limitations: As discussed at the beginning of Section IV, there are many sources of parameter variations in MOSFETs. For high-performance circuits, it is especially important to understand the effects of V_T variations, in particular. These variations may be either intradie or interdie and are caused by process variations, gate-length variations, process-induced proximity effects of neighboring structures, stochastic doping effects, and other intradie fluctuations [94]. These variations enter into consideration in three ways: 1) die-to-die and wafer-to-wafer variations in V_{T-CHIP}, the average V_T of a given chip; 2) an on-chip shift in I_{DDQ}, the total quiescent current of a die; and 3) V_{T-wc}, the worst case threshold voltage for any individual FET on a die.

The die-to-die variations in mean V_T can be characterized by an average value V_{T-nom} and a variance, although usually the worst case low average threshold V_{T-min} is of greatest interest since it is usually the fastest performing, both because of the low V_T and because low V_T is generally associated with shorter gates. Manufacturers sometimes take advantage of this spread by sorting the chips and selling the fastest ones at a higher price.

The intradie variations in V_T at the individual FET level are characterized by a Gaussian distribution $\rho(V_T)$ with standard deviation σ_{VT}. Using this distribution, one can calculate quiescent current

$$\begin{aligned} I_{DDQ} &= N\int_{V_{T\,min}}^{V_{T\,max}} \rho(V_T) I_{off}(V_T)\, dV_T \\ &= N\int \frac{1}{\sqrt{2\pi}\sigma_{VT}}\, e^{-(V_T - V_{T-CHIP})^2/2\sigma_{VT}^2} \\ &\quad \cdot I_{VT} e^{-V_T/\ln(10)S}\, dV_T \\ &= N I_{off-mean} e^{+\sigma_{VT}^2/2(\ln(10)S)^2} \end{aligned}$$

where N is the number of FETs and $NI_{off-mean}$ is the quiescent current that would flow if all the FETs on the entire chip had exactly the same threshold V_{T-CHIP}. Thus, it can be seen that I_{DDQ} is raised by the factor $e^{+\sigma_{VT}^2/10.58S^2}$ due to the variations in V_T.

In addition to its obvious use in determining whether a chip satisfies the leakage requirements of its particular application, I_{DDQ} also plays an important role in reliability strategies. Since I_{DDQ} is sensitive to several types of defects, including some that may not prevent functionality, it is used as a reliability criterion in many lower cost products. A measured low value indicates an absence of shorts and decreases the likelihood of latent defects in shipped products that may fail later in the field. In less cost-sensitive arenas many products are "burned in" by operating the product for a limited time at elevated voltages and/or temperatures. Both subthreshold conduction and oxide leakage are enhanced during such burn-in conditions, resulting in very high quiescent current.

Finally, V_{T-wc} the worst case threshold voltage for any individual MOSFET, is limited as well in at least two ways. First, digital CMOS circuits require adequate noise immunity to function properly. This constraint should scale at least approximately with V_{DD} and as a worst case, it has been estimated that V_{T-wc} must be at least 10% of V_{DD} based on analysis of 2.5- and 3.3-V logic technology [83]. Nominal V_T design targets are usually around 25% of V_{DD}. Second, some classes of high-performance circuits, such as dynamic logic, obtain performance advantage from nFET-dominated switching (nFET evaluate trees) and use very narrow pFETs or just node capacitance to hold the output high for a clock cycle prior to evaluation of the logic. If the V_T of a single nFET in such a position were to become too low, then subthreshold leakage would make the circuit nonfunctional. Since there are a very large number of devices on a chip, the statistics suggest that one should design to at least 6σ tolerances unless it can be shown that the V_T distribution falls off faster than Gaussian for low V_Ts.

2) High Power (30–100-W/cm^2 Active Power): For high-power applications, CV^2f is large, possibly reaching as high as 1 KW/cm^2 in high-activity factor macros at end-of-scaling logic density. As a result, under normal operating conditions the subthreshold leakage and gate tunneling leakage may be more limited by functionality concerns than by leakage power constraints. Because switching power is dominant, these designs do drive V_{DD} toward lower values which, in turn, drives V_T down. On the

other hand, the use of dynamic logic places an upper bound on individual FET leakage, creating a floor on V_T values and forcing the use of somewhat larger thresholds than would be allowed by power considerations alone. Use of multiple threshold values [95] can moderate this constraint, but does not entirely remove the problem. A high degree of reliability is usually required by systems employing such technologies and often demands special screening and/or burn-in procedures. Voltages from 25% to 100% above the nominal operating supply voltage and temperatures as high as 160 °C have often been employed. If, as is often the case, circuit functionality is required at these elevated temperatures and voltages, then the V_T floor necessary for dynamic logic and/or noise margins is raised significantly, especially compared to the low V_Ts suitable for chilled operation. Furthermore, in the limit the device cannot be scaled as far if it must sustain these higher voltages. Currently, the V_T situation is accommodated by using a higher V_T than desirable at operation conditions, solely so that burn-in conditions can be met. In future generations of technology, this burn-in constraint could be met through use of body-bias (or back-gate V_T control in DG-FETs) to dynamically increase the V_T during burn-in. In this case, it may be possible to push the nominal operating condition leakage currents up closer to the logic functionality limit, enabling lower threshold and supply voltages. In the end, reliability testing will force high-reliability parts to be larger and slower than "unreliable" parts that do not require such screening, and if V_T control is not used, they will be even slower due to the need for higher V_Ts.

The past scaling trend in this arena has been to scale V_{T-nom} with V_{DD} to maintain performance, while only allowing logarithmic scaling of V_{T-min} to contain increases in the worst case leakage currents. Clearly this results in a steadily decreasing V_T tolerance budget, which is getting more and more difficult to satisfy.

3) Medium-High Power (5–30 W/cm^2): This includes high-end desktop and midrange workstation processors. For these applications, leakage current can be limited by both power and functionality concerns, depending on design. Typically, leakage dissipation must be limited to some reasonable fraction of the total power budget at normal operation conditions and at this level, it will not impact functionality. Already some recent technologies with leakages in the realm of 10 nA/μm are beginning to experience this power limit. Reliability assurance will continue to be a burden, however, as in the high-power sector. These applications may also have low enough V_Ts that functionality is impacted during burn-in, possibly necessitating dynamically adjustable thresholds here, too. While the reliability requirements are typically less demanding, costs associated with the ability to burn in high I_{DDQ} designs may be prohibitive and, thus, drive lower leakage constraints on V_T and t_{ox} than the high-power sector. If dual V_T strategies are used, the low V_T may be limited to $\sim$180 mV at 25 °C based on the use of I_{DDQ} for reliability measurements. Similarly, t_{ox} may be constrained to 1.2-1.5 nm of silicon dioxide with a dual t_{ox} approach also likely to appear in order to better manage leakages.

4) Moderate Power (0.5–5 W/cm^2): Mobile processors, high-speed SRAM, and high-performance application-specific integrated circuits (ASICs) fall within this category. I_{DDQ} is limited simply due to power restrictions of the application. Loss of functionality due to leakage currents is not generally a concern. The low-power end of this range already has V_T limited to $\sim$300 mV for 0.1-μm MOSFETs and these thresholds will have to increase as scaling continues. Since quiescent power requirements are often much lower than the optimum active mode leakage power, gate oxides may be limited to a thicker range (1.7–2.0 nm) than would be possible for active mode optimization alone (1.3–1.6 nm). This application area is already experiencing smaller returns in performance with V_{DD} scaling than obtained in the past as a result of these pressures. Various well-biasing and power-supply switching schemes have been proposed as a means of allowing lower V_T for further performance leverage, but it is presently not clear just how much relief is practical. Techniques that either modify V_T dynamically or power V_{DD} locally must also be able to respond with very little delay. As a result, many solutions for the low-power arena (see Section V-D) may not be practical here.

VI. Discussion

Having described in some detail many of the physical effects that limit scaling of CMOS and the specific details of these limits for various applications, we now attempt to distill all of this information into a single table, Table 4. This table is intended to show the general landscape more than exact values and is compiled in the same spirit as Table 3: power density is the overriding parameter and the leakage mechanisms are each allocated as a certain fraction of the total power. More detailed optimizations might well change these fractions somewhat, but the overall trends described in the table should still be valid.

Table 4 is oriented around the power dissipation for active circuits. This focus is based on the assumption that quiescent power dissipation requirements during periods of long inactivity are best met by switching off the power supply. For applications for which this is not possible, it will be necessary to use higher thresholds, thicker oxides, and less aggressive doping than their active power limits would permit. As is the current practice, it is expected that multiple technologies may be present on the same chip, each optimized for its own particular power density target. Nevertheless, the assumption that the active power can be arbitrarily set to whatever limit the application demands needs further examination. The high-performance logic proposed in this table should easily be able to dissipate dynamic power at a rate of 1000 W/cm^2 if run at full speed with a fairly high activity factor (as is common for much of the logic in a high-performance processor). This is already ten times higher than one is probably willing to dissipate at the package. By lowering the speed somewhat, moving toward narrower longer devices at

Table 4
Estimated Scaling Limits for Various MOSFET Design Parameters and their Dependence on Application Class and Device Type

Device type	Application	T (°C)	Power (W/cm^2)	V_{DD} (V)	I_{off}^{max} (nA/μm)	S (mV/dec)	V_{Tn}^{min} (mV)	t_{oxTeq}^{min} (nm)	t_{Si}^{min} (nm)	Λ_{min} (nm)	L_{nom} (nm)
Bulk	High Performance	85	1000-30	0.8-1.2	1000[a]-110	99	140-235	1.0-1.3	8.5-11	9.5-13	14.5-19
		-40	1000-30	0.7-1.0	1000[a]-115	65	95-150	1.0-1.2	8-10	9-12	13.5-18
		-170	1000-30	0.5	1000[a]-155	33	50-75	0.9-1.1	8	9-9.5	14[b]
	....burn-in limited	140	-	1.8	1000[a]	115	180[c]	1.3	13.5	15	23
Bulk	Medium-High Performance	85	30-5	0.8-1.2	120-20	99	235-300	1.2-1.5	10-14	11.5-16	17-24
Bulk	Moderate Performance	85	5-0.5	0.6-1.0	25-2	99	300-390	1.3-1.6	10-14	11-16	17-24
Bulk	Low Power	65	0.5-0.001	0.7-0.9	1.0-0.01	94	410-550	1.7-2.0	13-17	15-19	22-29
Bulk	Ultra-Low Power	40	< 0.001	0.7-1.0	< 0.008	87	550-710	2.1-2.6	16-22	18-25	27-38
Bulk	Moderate Perf. SRAM	85	5-1	0.9-1.2	20-4	99	300-360	1.4-1.6	12-15	13.5-17	20-26
	Low Power SRAM	65	0.1-0.01	0.9-1.2	0.5-0.05	94	425-510	1.7-2.0	15-19	17-21.5	25-32
	Ultra-low power SRAM	40	0.0001	1.2	0.0006	87	635	2.4	23	26	39
Bulk	DRAM - metal gate	85	-	1.0	0.0001	99	790	2.5	28	31	49
Bulk	DRAM - neg. WL	85	-	1.0	0.0001	99	250	2.5	28	31	49
DG-FET	High Performance	85	1000-30	0.8-1.2	1000[a]-75	85	155-255	1.0-1.3	5-6	8.7	13[b]
		-40	1000-30	0.7-1.0	1000[a]-85	55	100-160	1.0-1.3	5-6	8.7	13[b]
		-170	1000-30	0.5	1000[a]-155	25	45-65	1.0-1.2	6-7	9.3	14[b]
DG-FET	Medium-High Performance	85	30-5	0.8-1.2	90-15	85	245-305	1.3-1.6	5-7	9-11.5	13[b]-17
DG-FET	Moderate Performance	85	5-0.5	0.6-1.0	20-2	85	300-390	1.3-1.7	5-6	9-10.5	13[b]-16
DG-FET	Low Power	65	0.5-0.001	0.7-0.9	0.7-0.007	80	420-510	1.7-2.1	5-8	9.5-13	14-20
DG-FET	Ultra-Low Power	40	< 0.001	0.7	< 0.005	75	530-660	2.1-2.5	5-8	10.5-15	16-22
		40	< 0.001	1.0	< 0.005	75	515-645	2.2-2.6	11-15	16-22	25-33
DG-FET	Moderate Perf. SRAM	85	5-1	0.9-1.2	14-3	85	315-355	1.5-1.7	5-9	9-13	13-20
	Low Power SRAM	65	0.1-0.01	0.9-1.2	0.3-0.04	80	425-475	1.8-2.1	6-13	10-18	15-27
	Ultra-low power SRAM	40	0.0001	1.2	0.0006	75	570	2.5	17	24	36

$I_{\rm off}^{\rm max}$, $\rm V_{Tn}^{min}$ and $\rm t_{oxTeq}^{min}$ are worst case values. Low temperature (-170C) bulk cases assume a forward body bias of 0.5 V, while all of the other bulk cases assume no forward body bias. PFET $|V_{Tp}^{\rm min}|$ is expected to be $\sim$50 mV lower than $\rm V_{Tn}^{min}$. To facilitate comparisons between cases, more precision is shown in some cases than is warranted on an absolute scale. Parameter ranges are intended to span the range of requirements and limits that might exist within the different application classes and are all organized in the same sense (from most aggressive scaling to least) with power and $\rm V_{DD}$ being independent variables. [a] indicates $I_{\rm off}^{\rm max}$ is limited by functionality concerns rather than by leakage dissipation. [b] indicates $L_{\rm nom}$ is limited by source-to-drain tunneling rather than by scale length. [c] indicates $\rm V_{Tn}^{min}$ is limited by noise margin requirements (10% of $\rm V_{DD}$) rather than by OFF current.

the lower power end of the table, and optimizing the thresholds and supplies for "low power," one may be able to lower the dynamic energy consumption per switching event by perhaps an order of magnitude. The speed of these technologies probably also decreases about an order of magnitude from the high-performance $\rm V_{DD} - V_T \simeq 1.0$ V cases to the $\rm V_{DD} - V_T \simeq 0.2$ V longer channel low-power cases. So, the overall situation is that the energy per logic operation only varies by about one order of magnitude over this table and the speed varies another order of magnitude, while the desired power varies by over six orders of magnitude. So how will the power density be lowered to the levels specified in the table? There are several approaches: 1) chips hardly ever use all of their circuitry so actively (by averaging over the less active areas and over large areas of lower dissipation SRAM or DRAM, one can probably reduce the power density an order of magnitude); 2) one can lower the clock frequency until one's throughput requirements are only just satisfied (this may also enable a further reduction in $\rm V_{DD}$, although $\rm V_{DD}$ cannot get too close to $\rm V_T$ because threshold variations cause too much timing uncertainty); 3) one can run the chip in bursts of power-optimized activity and turn it off between bursts; and 4) one can design the chip as many special purpose macros, each power- or energy-optimized for its specific task. The chip would then shuffle its work among the macros, minimizing the energy consumed, and increasing the averaging used in 1).

The methodology that has gone into creating Table 4 is as follows. Under worst case conditions (shortest gate length, lowest $\rm V_T$, thinnest insulator, highest temperature) 20% of the total power is allocated to subthreshold leakage currents, 10% is allocated to oxide tunneling currents, and 5% is allocated to band-to-band tunneling through the body-to-drain junction. These somewhat arbitrary percentages are chosen with the overall intent of allocating $\sim$2/3 of the power to useful switching, as in Section V-A. Since the final scaling limits are only logarithmically dependent on these percentages, the exact values are not critical. If $\rm V_{DD} \leq E_G/e - V_T^{nom}/\eta$, we assume there is no body-to-drain tunneling because the bands do not line up. For definiteness, we have taken $\rm V_T^{nom} = V_T^{min} + 50$ mV. The supply voltages do not represent scaling minima, but rather are simply estimates of

what we think will constitute reasonable supply levels for obtaining good performance. The supply voltage rises for ultralow-power applications simply to maintain some performance in the face of rising threshold voltages. The subthreshold slope is obtained from the given temperature and an assumed ideality factor at shortest gate length of 1.4 for the nonbody-biased bulk cases, 1.6 for the forward-body-biased bulk cases (those at −170°C), and 1.2 for the DG-FETs. The preceding are the only places where temperature enters: the tunneling processes are assumed to be temperature-independent.

The remaining variables are calculated self consistently using a spreadsheet. Beginning at the end, the nominal channel length is used to derive a total device width $W_{\rm tot}$ per cm^2 of Si, assuming that gates occupy 3% (estimated from ITRS99 [4]) of the total Si area and that the physical gate length is 40% longer than the effective channel length $L_{\rm nom}$. Assuming that on average half of the FETs are in the OFF state and contributing leakage current, the allocated subthreshold power is divided by the supply voltage and $W_{\rm tot}/2$ to obtain $I_{\rm off}$. The OFF current is, however, limited to a maximum of 1000 nA/μm as a rough estimate of the circuit functionality constraint. If the functionality constraint is reached, the gate and body-to-drain tunneling currents are also limited to 500 and 250 nA/μm, respectively. The nFET threshold voltage $\rm V_{Tn}$ is then obtained from $\rm V_{Tn} = S\log_{10}(I_0/L_{nom}I_{off})$, where I_0 is 300 nA for bulk devices and 600 nA for DG-FETs (since they effectively have twice the width). The maximum gate insulator tunneling current density is determined by dividing the power allocated to gate tunneling by the supply voltage and half the total gate area, again on the assumption that only half the gates are in the turned-on state. DG-FETs use 6% for the total gate area since there are two surfaces, which are assumed symmetric for this analysis. This tunneling current density is translated into a minimum equivalent tunneling oxide thickness $\rm t_{oxTeq}^{min}$ using the curves in Fig. 7. We are optimistically assuming here that oxide/insulator reliability problems can be solved to the extent that they will not impose lower current density limits for the supply voltages considered. If these problems cannot be solved, the high-performance cases might need substantially thicker insulators, possibly increasing their channel lengths 10%–20%.

The minimum scale length $\Lambda_{\rm min}$ is derived from the body-to-drain tunneling constraint. For the 25-nm bulk design the vertical extent of the body-to-drain tunneling region is ∼10–15 nm, which is $\sim L/2$ (see Fig. 15). We have assumed that this dimension will scale with channel length (and that half the FETs are in the OFF state) and so the power allocated to this tunneling is divided by the supply voltage and $W_{\rm tot}/2$ times $L_{\rm nom}/2$ to obtain a maximum tunneling current density, which is converted into a maximum field $F_{\rm max}$ using the curve in Fig. 9. $\Lambda_{\rm min}$ is then obtained from this field by scaling the 25-nm design point: $\Lambda_{\rm min} = \Lambda_{25\ \rm nm}(1.75\ \rm MV/cm/F_{max})((V_{DB} + E_G/e)/2.1\ V)$, where $\Lambda_{25\ \rm nm} = 16.7$ nm, and $\rm V_{DB} = V_{DS} - V_{BS}$ is the drain-to-body voltage. The calculation for DG-FETs is similar except the cross-sectional area for tunneling uses the Si film thickness instead of half the channel length. In this case, $\Lambda_{\rm min}$ is determined by scaling from a $\Lambda_1 = 10.2$ nm DG-FET simulation as a function of $\rm V_{DS} + V_T/\eta$.

Next, $\rm t_{Si}^{min}$ is evaluated using the preceding $\Lambda_{\rm min}$, $\rm t_{oxTeq}^{min}$ and the scale length theory presented in Section II-A and the Appendix. Since these are supposed to be limits to scaling, we assumed that the gate insulator is like Al_2O_3 with a tunneling barrier similar to SiO_2, but twice the dielectric constant. If alternate dielectric materials cannot be used and FETs are stuck with oxy-nitrites, then the channel lengths in Table 4 would need to be increased 0%–20% depending on the extent to which the increased insulator thickness can be countered by thinning $\rm t_{Si}^{min}$. For DG-FETs, if the computed $\rm t_{Si}^{min}$ is smaller than the minimum Si layer thickness (5 nm, based on the tolerance-related principles discussed in Section IV-B) or if it does not exist because $\rm V_{DD} \le E_G/e - V_T^{nom}/\eta$, then $\rm t_{Si}^{min}$ is set to 5 nm and $\Lambda_{\rm min}$ is recomputed from $\rm t_{Si}^{min}$ and $\rm t_{oxTeq}^{min}$. For the cases in which source-to-drain tunneling constrains the value of $L_{\rm nom}$, we set $\rm t_{Si}^{min}$ to yield the constrained value for $L_{\rm nom}$ since there is no point in making it smaller.

Finally, the nominal channel lengths $L_{\rm nom}$ are determined by multiplying the scale length by 1.5 for logic MOSFETs and 1.6 for the DRAM. This assumes that both bulk and DG-FETs have some sort of compensation for $\rm V_T$ rolloff (e.g., halos), although the feasibility of such compensation for the smallest devices is certainly questionable.

This table reveals the dependence of scaling limits on applications very clearly. As one moves from high- to low-power applications, the shrinking leakage requirements cause the minimum allowed nominal channel length for bulk MOSFETs to increase from ∼14 nm to ∼38 nm, almost 3×, while $\rm t_{oxTeq}^{min}$ increases from 1 to 2.6 nm. DG-FETs channel lengths also increase from 13 to 33 nm, revealing a 10%–50% scaling advantage depending on conditions. The advantage is low for the shortest channel cases because both devices are limited by source-to-drain tunneling and is also low for the high $\rm V_T$, high $\rm V_{DD}$ cases where the DG-FET is more impacted by body-to-drain tunneling. For the majority of applications, however, especially those at lower $\rm V_{DD}$, the DG-FET shows a very substantial advantage, equivalent to an entire generation of scaling. The $\rm V_{DD}$ dependence is particularly evident in the ultralow-power case. Note that the DG-FETs tend to have lower threshold voltages for the low-power cases because of their steeper subthreshold slopes, but this is partly compensated by decreasing $I_{\rm off}$ requirements due to their greater density. On the performance side, the DG-FETs can probably significantly outperform their bulk counterparts in wiring capacitance-dominated circuits because of their effectively doubled current drive (two gated surfaces). Overall, for both device types, traditional $\rm V_T/V_{DD}$ ratios only appear to be achievable for the highest performance design point. The lower power points require progressively higher ratios than have been used historically.

The temperature dependence of the high-performance design points shows that although temperature may not buy any advantage as far as minimum channel length is concerned, it should allow lower voltage operation. This is

very important since lower voltage means lower energy per logic operation and, hence, more operations can be done per second for the same total chip power dissipation; i.e., the clock can be sped up or the logic made more dense. The low temperatures should also enhance device speed by improving mobility. The −170 °C 0.5-V design points are interesting because their nominal design points are limited by source-to-drain tunneling rather than by their scale lengths. This is because source-to-drain tunneling interferes with the steep low-temperature subthreshold slopes a little sooner than it strikes the less-steep room temperature curves (see Section III). The high-temperature burn-in limited case shows that the scale length must be increased significantly to satisfy the high burn-in voltage requirement. Also, V_T must be set quite high to meet margin requirements well above the desired V_T for the lower temperature operating conditions, resulting in further reduced performance if a dynamically adjustable V_T approach (e.g., body-biasing) is not adopted.

The DRAM design points are based on the criteria in Section V-B and show two possibilities with regard to achieving the necessary OFF current. For the wordline (WL) not going below ground, a midgap workfunction metal gate would probably be necessary to achieve the desired threshold. For a negative wordline-low voltage, however, a more reasonable logic-like threshold could be used, as described in Section V-B.

The bulk MOSFET projections in Table 4 all assume that the body is at a fixed bias of zero, except for the −170 °C cases, which assume 0.5-V forward bias. For PD-SOI MOSFETs, this is not usually the case. The high drain-to-body tunneling currents assumed in these extremely scaled FETs would forward bias the floating body much further than occurs in current designs. From the point of view of scale length, this would seem advantageous because like the −170 °C case, it would reduce the depletion depth, which would enable further scaling. Another effect is that the forward body bias may be strongly dependent on the drain voltage, potentially causing large output conductance such as that seen recently in 52-nm physical gate length PD-SOI [96]. If V_{DD} is increased on the DG-FET designs, the same sort of effect would be expected. More research is needed into this regime to determine the relative advantages and disadvantages of forward body bias and high output conductance in SOI.

These scaling limit projections rest mostly on the leakage current mechanisms discussed earlier in this paper. How accurate are they? The required threshold voltages should be reasonably accurate since they depend only on the V_T definition itself and the well-understood dependence of the subthreshold current on kT and ideality. Only for the high-performance DG-FET and −170°C cases does the subthreshold enter into the less-well characterized source-to-drain tunneling regime. The t_{oxTeq}^{min} requirements are based on oxide tunneling curves that have been well measured in recent years. Although there is some disagreement about how to determine the physical thickness of very thin oxide layers, t_{oxTeq}^{min} should really be thought of as a parameterization of the current density that can be applied to any insulator. Unfortunately, the most sensitive parameter in determining the minimum scaling dimension is the depletion depth, which is determined from the band-to-band tunneling curve in Fig. 9. This curve admittedly rests on relatively little data and deserves much further investigation because it plays such a prominent role in the end of scaling.

For the DG-FET limits, one big uncertainty about these projections is the question of whether high channel mobility can indeed be obtained since this has not yet been successfully demonstrated experimentally. Also, obtaining the desired V_Ts is difficult for DG-FETs since the channel is not thick enough to allow V_T shifts of more than about ±100 mV by use of doping without causing excessive V_T fluctuations due to the discreteness of the dopants [54]. Indeed, it would be better not to put any dopants in the channel at all. Consequently, the DG-FET design points probably require the development of several different metal gate technologies with suitable workfunctions. Drain-to-floating-body tunneling in DG-FETs is also an issue and needs further work.

Although the designs presented in the table account for ∼25% lithographic variation in channel length by setting the nominal channel longer than the minimum channel length, they do not explicitly take into account threshold variations due to dopant fluctuations. As pointed out in Section IV-C, these variations are already substantial for the 25-nm MOSFET design and will only grow larger in scaling to the 15-nm design point suggested in the table. This problem primarily afflicts the bulk designs (if the DG-FET V_Ts are obtained by workfunction adjustment) and will necessitate raising the V_T targets—especially for SRAM. This effect appears to give an added advantage to DG-FETs.

For lack of expertise, we have not discussed analog device application constraints in this paper, but a few general comments are possible. Two of the main requirements of analog applications are "higher" supply voltages (1.5–2.0 V often seems to be the minimum acceptable) and high-output resistance. The higher voltage will necessitate thicker oxide, 2–3 nm depending perhaps more on reliability concerns than on leakage current. The high-output resistance translates into a large L/Λ_1 ratio according to Fig. 5, perhaps 3.0 or more. Based on these assumptions, it appears that the scaling limit for generic analog applications is considerably larger than for logic, probably in the ∼80-nm channel length regime.

VII. Conclusion

We have described most of the important physical phenomena that stand in the way of continued scaling of Si CMOS technology and have shown how these effects determine different limits for different circuit applications. Most of the application limits are set by limitations on the amount of power that can be dissipated in the three primary leakages: subthreshold channel current, gate-to-channel tunneling through the insulator, and body-to-drain junction tunneling currents. Source-to-drain tunneling along the channel is also a possible limitation for very short channels and at low temperature.

The scale length theory that has been presented here provides a useful framework within which to understand the tradeoff between channel length and short channel effects. Using this theory in conjunction with the various limiting effects, we have projected that bulk-like CMOS should be extendible down to about 14-nm nominal channel length for high-performance logic and ~35 nm for very low power applications, with intermediate applications falling in between. DG-FETs are projected to be scalable to 10%–50% shorter channel lengths than their bulk counterparts with the greatest advantage being at low-supply voltages. These estimates include allowance for reasonable tolerances between nominal and worst case channel lengths as required for current manufacturing processes, but for experiments in which tolerances can be ignored we have discussed several results for FETs with channel length in the 8–12-nm range.

Overall, we conclude that there is no single endpoint to scaling of CMOS. Rather, there are many endpoints, each optimally adapted to its particular applications. As the industry moves forward, greater flexibility will have to be developed in manufacturing different devices for different users.

APPENDIX

Following the method of [19], consider the idealized MOSFET cross section in Fig. 3(a), which defines x, y, the depletion depth $\mathrm{t_{Si}}$, and the dielectric thickness t_I. Using superposition, the potentials ψ_1 and ψ_2 in the center of the FET can be written as

$$\begin{aligned}\psi_1 &= u_{L1}(x,y) + u_{R1}(x,y) + v_1(x) + v_{D1}(x,y)\\ \psi_2 &= u_{L2}(x,y) + u_{R2}(x,y) + v_2(x) + v_{D2}(x,y)\end{aligned} \tag{A1}$$

where the v_{Di} satisfy Poisson's equation for the fixed charges (usually nonuniform) in the device with all boundary potentials at zero, v_i are the one-dimensional solutions to Poisson's equation satisfying the gate voltage, body voltage, and dielectric boundary conditions, and u_{Li} and u_{Ri} are left and right $\sinh y \sin x$ solutions to Laplace's equation, which accommodate potentials applied to the source and drain [21].

These us can be written as infinite series, but the lowest order term is the most important, since the higher terms decay very rapidly

$$\begin{aligned}u_{L1} &\simeq b_{11}\frac{\sinh(k_1(L-y))}{\sinh(k_1 L)}\sin(k_1(x+\mathrm{t}_I))\\ u_{R1} &\simeq c_{11}\frac{\sinh(k_1 y)}{\sinh(k_1 L)}\sin(k_1(x+\mathrm{t}_I))\\ u_{L2} &\simeq b_{21}\frac{\sinh(k_1(L-y))}{\sinh(k_1 L)}\sin(k_1(x-\mathrm{t_{Si}})+\pi)\\ u_{R2} &\simeq c_{21}\frac{\sinh(k_1 y)}{\sinh(k_1 L)}\sin(k_1(x-\mathrm{t_{Si}})+\pi)\end{aligned} \tag{A2}$$

where the bs, cs, and $k_1 = \pi/\Lambda_1$ are coefficients to be determined by satisfying the boundary conditions. These forms are chosen to provide $u = 0$ at the $x = -\mathrm{t}_I$ and $x = \mathrm{t_{Si}}$ boundaries and one half period of the sine functions in between. As shown in Fig. 4, these analytic solutions very accurately approximate the potential variation along the channel under the gate.

By constraining the upper and lower analytic solutions to satisfy the usual dielectric boundary conditions at $x = 0$, one can arrive at an equation for the eigenvalues, of which k_1 is the smallest

$$0 = \epsilon_{\mathrm{Si}}\tan(k_1\mathrm{t}_I) + \epsilon_I\tan(k_1\mathrm{t_{Si}}) \tag{A3}$$

or, in terms of the scale length Λ_1

$$0 = \epsilon_{\mathrm{Si}}\tan(\pi\mathrm{t}_I/\Lambda_1) + \epsilon_I\tan(\pi\mathrm{t_{Si}}/\Lambda_1). \tag{A4}$$

In a similar manner, a device containing three layers of dielectrics with thicknesses t_1, t_2, and t_3 and dielectric constants ϵ_1, ϵ_2, and ϵ_3 can be solved for an eigenvalue equation for the lowest order scale length satisfying both sets of dielectric boundary conditions

$$\begin{aligned}&\frac{\epsilon_2}{\epsilon_1\epsilon_3}\tan(k_1\mathrm{t}_1)\tan(k_1\mathrm{t}_2)\tan(k_1\mathrm{t}_3)\\ &\quad= \frac{1}{\epsilon_1}\tan(k_1\mathrm{t}_1) + \frac{1}{\epsilon_2}\tan(k_1\mathrm{t}_2) + \frac{1}{\epsilon_3}\tan(k_1\mathrm{t}_3).\end{aligned} \tag{A5}$$

This solution could be used to analyze a bulk MOSFET with a two-layer insulator (e.g., a thin SiO_2 layer followed by a high-k layer) or it can be applied to a DG-FET in which case the center layer is silicon. If the DG-FET is symmetric, as in Fig. 3(b), the eigenvalue equation can be simplified to

$$1 = \frac{\epsilon_{\mathrm{Si}}}{\epsilon_I}\tan(\pi\mathrm{t}_I/\Lambda_1)\tan(\pi\mathrm{t_{Si}}/2\Lambda_1). \tag{A6}$$

By differentiating this equation with respect to $\mathrm{t_{Si}}$, one can arrive at an expression analogous to (2) relating $\Delta\mathrm{t_{Si}}$ to ΔL at constant L/Λ_1

$$\frac{\Delta L}{L} = \frac{\mathrm{sinc}(2\pi\mathrm{t}_I/\Lambda_1)}{\mathrm{sinc}(\pi\mathrm{t_{Si}}/\Lambda_1) + \mathrm{sinc}(2\pi\mathrm{t}_I/\Lambda_1)}\cdot\frac{\Delta\mathrm{t_{Si}}}{\mathrm{t_{Si}}} \tag{A7}$$

where the function $\mathrm{sinc}(u) = \sin(u)/u$.

ACKNOWLEDGMENT

The authors would like to thank the invaluable contributions of many of their colleagues, including E. Jones, L. Huang, and G. Cohen for their collaboration on DG-FET research, M. Ieong for his work on device modeling, J. Stathis for consultations about oxide reliability, and J. Welser and S. Schuster for useful discussions. The authors would also like to thank Y. Naveh and K. Likharev for useful discussions about the limits of DG-FET scaling and for access to several of their preprints and H. Kawaura for kindly providing original versions of his figures. The management support of J. Warlaumont is also greatly appreciated.

REFERENCES

[1] J. E. Lilienfeld, "Method and apparatus for controlling electric currents," U.S. Patent 1 745 175, 1930.

[2] D. Kahng and M. M. Atalla, "Silicon–silicon dioxide field induced surface devices," presented at the IRE Solid-State Device Res. Conf., Pittsburgh, PA, June 1960.

[3] P. K. Bondy, "Moore's law governs the silicon revolution," *Proc. IEEE*, vol. 86, pp. 78–81, Jan. 1998.

[4] Semiconductor Industry Association (SIA), *International Technology Roadmap for Semiconductors*, 1999 ed. San Jose, CA: SIA, 1999.

[5] B. Davari, R. H. Dennard, and G. G. Shahidi, "CMOS scaling for high-performance and low-power—the next ten years," *Proc. IEEE*, vol. 89, pp. 595–606, Apr. 1995.

[6] Y. Taur, D. Buchanan, W. Chen, D. Frank, K. Ismail, S.-H. Lo, G. Sai-Halasz, 6R. Viswanathan, H.-J. C. Wann, S. Wind, and H.-S. Wong, "CMOS scaling into the nanometer regime," *Proc. IEEE*, vol. 85, pp. 486–504, Apr. 1997.
[7] S. Asai and Y. Wada, "Technology challenges for integration near and below 0.1 μm," *Proc. IEEE*, vol. 85, pp. 505–520, Apr. 1997.
[8] T. Sugii, Y. Momiyama, M. Deura, and K. Goto, "MOS scaling beyond 0.1 μm," in *Silicon Nanoelectronics Workshop*, June 1999, pp. 60–61.
[9] H.-S. P. Wong, D. J. Frank, P. M. Solomon, H.-J. Wann, and J. Welser, "Nanoscale CMOS," *Proc. IEEE*, vol. 87, pp. 537–570, Apr. 1999.
[10] R. Yan, A. Ourmazd, and K. F. Lee, "Scaling the Si MOSFET: From bulk to SOI to bulk," *IEEE Trans. Electron Devices*, vol. 39, pp. 1704–1710, July 1992.
[11] 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. Dig.*, pp. 553–556, 1992.
[12] 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 MOSFETs at the 25 nm channel length generation," *IEDM Tech. Dig.*, p. 407, 1998.
[13] T. Ernst, C. Tinella, C. Raynaud, and S. Cristoloveanu, "Fringing fields in sub-0.1 μm FD SOI MOSFETs: Optimization of the device architecture," in *Proc. ULIS 2000 Workshop* , Jan. 2000, pp. 47–50.
[14] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 256–268, Oct. 1974.
[15] D. L. Critchlow, "MOSFET scaling—The driver of VLSI technology," *Proc. IEEE*, vol. 87, pp. 659–667, Apr. 1999.
[16] Y. Taur and E. Nowak, "CMOS devices below 0.1 μm: How high will performance go?," in *IEDM Tech. Dig.*, 1997, pp. 215–218.
[17] 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*, vol. ED-31, pp. 452–462, Apr. 1984.
[18] D. J. Frank, "Application and technology forecast," in *Low Power Design in Deep Submicron Electronics*, W. Nebel and J. Mermet, Eds. Norwell, MA: Kluwer, 1997, vol. 337, pp. 9–44.
[19] D. J. Frank, Y. Taur, and H.-S. P. Wong, "Generalized scale length for two-dimensional effects in MOSFETs," *IEEE Electron Device Lett.*, vol. 19, pp. 385–387, Oct. 1998.
[20] K. N. Ratnakumar and J. D. Meindl, "New IGFET short-channel threshold voltage model," in *IEDM Tech. Dig.*, 1981, pp. 204–206.
[21] T. N. Nguyen, "Small-geometry MOS transistors: Physics and modeling of surface- and buried-channel MOSFETs," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1984.
[22] R. M. Swanson and J. D. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE J. Solid-State Circuits*, vol. SC-7, pp. 146–153, Apr. 1972.
[23] T. Ghani, S. Ahmed, P. Aminzadeh, J. Bielefeld, P. Charvat, C. Chu, M. Harper, P. Jacob, C. Jan, J. Kavalieros, C. Kenyon, R. Nagisetty, P. Packan, J. Sebastian, M. Taylor, J. Tsai, S. Tyagi, S. Yang, and M. Bohr, "100 nm gate length high performance/low power CMOS transistor structure," in *IEDM Tech. Dig.*, 1999, pp. 415–418.
[24] S.-H. Lo, D. A. Buchanan, Y. Taur, and W. Wang, "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFETs," *IEEE Electron Device Lett.*, vol. 18, p. 209, May 1997.
[25] S.-H. Lo, D. A. Buchanan, and Y. Taur, "Modeling and characterization of n^+- and p^+-polysilicon-gated ultra thin oxides (21–26A)," in *Proc. Symp. VLSI Technol.*, June 1997, pp. 555–1212.
[26] Y. Taur, Y.-J. Mii, D. J. Frank, H.-S. Wong, D. A. Buchanan, S. J. Wind, S. A. Rishton, G. A. Sai-Halasz, and E. J. Nowak, "CMOS scaling into the 21st century: 0.1 μm and beyond," *IBM J. Res. Dev.*, vol. 39, p. 245, 1995.
[27] Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS design considerations," in *IEDM Tech. Dig.*, 1998, pp. 789–792.
[28] T. Ghani, K. Mistry, P. Packan, S. Thompson, M. Stettler, S. Tyagi, and M. Bohr, "Scaling challenges and device design requirements for high performance sub-50 nm gate length planar CMOS transistors," in *Proc. Symp. VLSI Technol.*, June 2000, pp. 174–175.
[29] H.-S. Wong, D. J. Frank, Y. Taur, and J. M. C. Stork, "Design and performance considerations for sub-0.1 μm double-gate SOI MOSFETs," in *IEDM Tech. Dig.*, 1994, pp. 747–750.
[30] C. P. Auth and J. D. Plummer, "Scaling theory for cylindrical fully-depleted, surrounding-gate MOSFETs," *IEEE Electron Device Lett.*, vol. 18, p. 74, Feb. 1997.
[31] S.-H. Oh, D. Monroe, and J. M. Hergenrother, "Analytic description of short-channel effects in fully-depleted double-gate and cylindrical, surrounding-gate MOSFETs," *IEEE Electron Device Lett.*, vol. 21, pp. 445–447, Sept. 2000.
[32] J. J. Welser, S. Tiwari, and P. M. Solomon, "Straddle-gate transistor: Changing MOSFET channel length between off- and on-state toward achieving tunneling-defined limit of field-effect," in *IEDM Tech. Dig.*, 1998, pp. 737–740.
[33] D. J. Frank and H.-S. P. Wong, "Analysis of the design space available for high-k gate dielectrics in nanoscale MOSFETs," in *Proc. Silicon Nanoelectronics Workshop*, June 2000, pp. 47–48.
[34] T. Hamamoto, S. Sugiura, and S. Sawada, "On the retention time distribution of dynamic random access memory (DRAM)," *IEEE Trans. Electron Devices*, vol. 45, pp. 1300–1309, June 1998.
[35] S. Kamohara, K. Kubota, M. Moniwa, K. Ohyu, and A. Ogishima, "Statistical PN junction leakage model with trap level fluctuation for Tref (refresh time)-oriented DRAM design," in *IEDM Tech. Dig.*, 1999, pp. 539–542.
[36] H. Kawaura, T. Sakamoto, and T. Baba, "Direct source–drain tunneling current in subthreshold region of sub-10-gate EJ-MOSFETs," in *Si Nanoelectronics Workshop Abstracts*, June 1999, pp. 26–27.
[37] C. Wann, J. Harrington, R. Mih, S. Biesemans, K. Han, R. Dennard, O. Prigge, C. Lin, and R. Mahnkopf, "CMOS with active well bias for low-power and RF/analog applications," in *Proc. Symp. VLSI Technol.*, June 2000, pp. 158–159.
[38] Y. Mii, S. Wind, Y. Taur, Y. Lii, D. Klaus, and J. Bucchignano, "An ultra-low power 0.1 μm CMOS," in *Proc. Symp. VLSI Technol.*, June 1994, pp. 9–10.
[39] M. Inohara, H. Oyamatsu, Y. Unno, Y. Fukaura, S. Goto, Y. Egi, and M. Kinugawa, "Highly scalable and fully logic compatible SRAM cell technology with metal damascene process and W local interconnect," in *Proc. Symp. VLSI Technol.*, June 1998, pp. 64–65.
[40] S. Deleonibus, C. Caillat, G. Guegan, M. Heitzmann, M. E. Nier, S. Tedesco, B. Dal'zotto, F. Martin, P. Mur, A. M. Papon, G. Lecarval, S. Biswas, and D. Souil, "A 20 nm physical gate length NMOSFET featuring 1.2 nm gate oxide, shallow implanted source and drain and BF2 pockets," *IEEE Electron Device Lett.*, vol. 47, pp. 173–175, Apr. 2000.
[41] F. Assad, Z. Ren, D. Vasileska, S. Datta, and M. Lundstrom, "On the performance limits for Si MOSFETs: A theoretical study," *IEEE Trans. Electron Devices*, vol. 47, pp. 232–240, Jan. 2000.
[42] F. G. Pikus and K. K. Likharev, "Nanoscale field-effect transistors: An ultimate size analysis," *Appl. Phys. Lett.*, vol. 71, no. 25, pp. 3661–3663, Dec. 1997.
[43] Y. Naveh and K. K. Likharev, "Modeling of 10-nm-scale ballistic MOSFETs," *IEEE Electron Device Lett.*, vol. 21, pp. 242–244, May 2000.
[44] D. Liu and C. Svensson, "Trading speed for low power by choice of supply and threshold voltages," *IEEE J. Solid-State Circuits*, vol. 28, pp. 10–17, Jan. 1993.
[45] Z. Chen, J. Burr, J. Shott, and J. D. Plummer, "Optimization of quarter micron MOSFETs for low voltage/low power applications," in *IEDM Tech. Dig.*, 1995, pp. 63–66.
[46] D. J. Frank, P. Solomon, S. Reynolds, and J. Shin, "Supply and threshold voltage optimization for low power design," in *Proc. Int. Symp. Low Power Electron. Design*, Aug. 1997, pp. 317–322.
[47] K. A. Bowman, X. Tang, J. C. Eble, and J. D. Meindl, "Impact of extrinsic and intrinsic parameter variations on CMOS system on a chip performance," in *Proc. 12th Annu. IEEE Int. ASIC/SOC Conf.*, Sept. 1999, pp. 267–271.
[48] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[49] D. J. Frank, Y. Taur, M. Ieong, and H.-S. P. Wong, "Monte Carlo modeling of threshold variation due to dopant fluctuations," in *Proc. Symp. VLSI Technol.*, June 1999, pp. 169–170.
[50] S. E. Laux, M. V. Fischetti, and D. J. Frank, "Monte Carlo analysis of semiconductor devices: The DAMOCLES program," *IBM J. Res. Dev.*, vol. 34, p. 466, July 1990.
[51] C. Fiegna, H. Iwai, T. Wada, T. Saito, E. Sangiorgi, and B. Ricco, "A new scaling methodology for the 0.1–0.025 μm MOSFET," in *Proc. Symp. VLSI Technol.*, June 1992, p. 33.
[52] K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, "Scaling theory for double-gate SOI MOSFETs," *IEEE Trans. Electron Devices*, vol. 40, p. 2326, Dec. 1993.

[53] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1956, ch. 11, p. 283.
[54] H.-S. P. Wong, "Novel device options for sub-100 nm CMOS," in *IEDM Short Course: Sub-100 nm CMOS*, M. Bohr, Ed. Piscataway, NJ: IEEE Press, 1999.
[55] M. Ieong and H.-S. P. Wong, "Analysis of 25 nm double-gate MOSFETs including self-consistent 2-D quantization effects," *IEEE Electron Device Lett.*, submitted for publication.
[56] M. V. Fischetti, "A master equation approach to the study of electronic transport in small semiconductor devices," *Phys. Rev. B*, vol. 59, no. 7, pp. 4901–4917, Feb. 1999.
[57] S. E. Laux and M. V. Fischetti, "Monte Carlo study of velocity overshoot in switching a 0.1-micron CMOS inverter," in *IEDM Tech. Dig.*, 1997, pp. 877–880.
[58] F. Gamiz, J. A. Lopez-Villanueva, J. B. Roldan, J. E. Carceller, and P. Cartujo, "Monte Carlo simulation of electron transport properties in extremely thin SOI MOSFETs," *IEEE Trans. Electron Devices*, vol. 45, pp. 1122–1126, May 1998.
[59] F. Gamiz, J. B. Roldan, P. Cartujo-Cassinello, J. E. Carceller, J. A. Lopez-Villanueva, and S. Rodriguez, "Electron mobility in extremely thin single-gate silicon-on-insulator inversion layers," *J. Appl. Phys.*, vol. 86, no. 11, pp. 6269–6275, 1999.
[60] M. Shoji and S. Horiguchi, "Electronic structures and phonon limited electron mobility of double-gate silicon-on-insulator Si inversion layers," *J. Appl. Phys.*, vol. 85, no. 5, pp. 2722–2731, 1999.
[61] A. Toriumi, J. Koga, H. Satake, and A. Ohata, "Performance and reliability concerns of ultra-thin SOI and ultra-thin gate oxide MOSFETs," in *IEDM Tech. Dig.*, 1995, pp. 847–850.
[62] J.-H. Choi, Y.-J. Park, and H.-S. Min, "Electron mobility behavior in extremely thin SOI MOSFETs," *IEEE Electron Device Lett.*, vol. 16, pp. 527–529, Nov. 1995.
[63] T. Ernst, D. Munteanu, S. Cristoloveanu, T. Ouisse, S. Horiguchi, Y. Ono, Y. Takahashi, and K. Murase, "Investigation of SOI MOSFETs with ultimate thickness," *Microelectron. Eng.*, vol. 48, pp. 339–342, June 1999.
[64] H.-S. Wong, K. Chan, and Y. Taur, "Self-aligned (top and bottom) double-gate MOSFET with a 25 nm thick silicon channel," in *IEDM Tech. Dig.*, 1997, pp. 427–430.
[65] D. Hisamoto, W.-C. Lee, J. Kedzierski, E. Anderson, H. Takeuchi, K. Asano, T.-J. King, J. Bokor, and C. Hu, "A folded-channel MOSFET for deep-sub-tenth micron era," in *IEDM Tech. Dig.*, 1998, pp. 1032–1034.
[66] X. Huang, W.-C. Lee, C. Ku, 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," in *IEDM Tech. Dig.*, 1999, pp. 67–70.
[67] J.-H. Lee, G. Tarashi, A. Wei, T. A. Langdo, E. A. Fitzgerald, and D. A. Antoniadis, "Super self-aligned double-gate (SSDG) MOSFETs utilizing oxidation rate difference and selective epitaxy," in *IEDM Tech. Dig.*, 1999, pp. 71–74.
[68] R. W. Keyes, "The effect of randomness in the distribution of impurity atoms on FET thresholds," *Appl. Phys.*, vol. 8, pp. 251–259, 1975.
[69] V. De, X. Tang, and J. Meindl, "Random MOSFET parameter fluctuation limits to gigascale integration (GSI)," in *Proc. Symp. VLSI Technol.*, June 1996, pp. 198–199.
[70] X. Tang, V. K. De, and J. D. Meindl, "Intrinsic MOSFET parameter fluctuations due to random dopant placement," *IEEE Trans. VLSI Syst.*, vol. 5, pp. 369–376, Dec. 1997.
[71] V. K. De, X. Tang, and J. D. Meindl, "Scaling limits of Si MOSFET technology imposed by random parameter fluctuations," in *Proc. IEEE Device Res. Conf. Dig.*, June 1996, pp. 114–115.
[72] Y. Yasuda, M. Takamiya, and T. Hiramoto, "Effects of impurity position distribution on threshold voltage fluctuations in scaled MOSFETs," in *Si Nanoelectronics Workshop Abstracts*, June 1999, pp. 26–27.
[73] H.-S. Wong and Y. Taur, "Three-dimensional 'atomistic' simulation of discrete microscopic random dopant distributions effects in sub-0.1 μm MOSFETs," in *IEDM Tech. Dig.*, 1993, pp. 705–708.
[74] H.-S. P. Wong, Y. Taur, and D. Frank, "Discrete random dopant distribution effects in nanometer-scale MOSFETs," *Microelectron. Reliability*, vol. 38, no. 9, pp. 1447–1456, 1998.
[75] A. Asenov and S. Saini, "Random dopant fluctuation resistant decanano MOSFET architectures," in *Si Nanoelectronics Workshop Abstracts*, June 1999, pp. 84–85.
[76] P. M. Solomon, "A comparison of semiconductor devices for high speed logic," *Proc. IEEE*, vol. 70, pp. 489–509, May 1982.
[77] A. Asenov, G. Slavcheva, A. R. Brown, J. H. Davies, and S. Saini, "Quantum mechanical enhancement of the random dopant induced threshold voltage fluctuations and lowering in sub 0.1 micron MOSFETs," in *IEDM Tech. Dig.*, 1999, pp. 535–538.
[78] D. Burnett, K. Erington, C. Subramanian, and K. Baker, "Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits," in *Proc. Symp. VLSI Technol.*, June 1994, pp. 15–16.
[79] D. J. Frank and H.-S. P. Wong, "Simulation of stochastic doping effects in Si MOSFETs," in *Proc. Int. Workshop Computational Electron.*, May 2000, pp. 2–3.
[80] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.*, vol. EDL-2, pp. 126–129, May 1981.
[81] D. J. Frank, "Design considerations for CMOS near the limits of scaling," in *Proc. ULIS 2000 Workshop*, Jan. 2000, pp. 3–7.
[82] J. H. Stathis, A. Vayshenker, P. R. Varekamp, E. Y. Wu, C. Montrose, J. McKenna, D. J. DiMaria, L.-K. Han, E. Cartier, R. A. Wachnik, and B. P. Linder, "Breakdown measurements of ultra-thin SiO_2 at low voltage," in *Proc. Symp. VLSI Technol.*, June 2000, pp. 94–95.
[83] R. H. Dennard, "Scaling challenges for DRAM and microprocessors in the 21st century," in *Proc. Int. Symp. ULSI Science Technol.*, 1997, pp. 519–532.
[84] M. Asakura *et al.*, "A 34 nm 256 Mb DRAM with boosted sense-ground scheme," in *Int. Solid State Circuits Conf., Dig. Tech. Papers*, 1994, pp. 140–141.
[85] J. Chen, T. Y. Chan, I. C. Chen, P. K. Ko, and C. Hu, "Sub-breakdown drain leakage current in MOSFET," *IEEE Electron Device Lett.*, vol. EDL-8, pp. 515–518, Nov. 1987.
[86] Gruening *et al.*, "A novel trench DRAM cell with a VERtIcal access transistor and BuriEd STrap (VERI BEST) for 4 Gb/16Gb," in *IEDM Tech. Dig.*, 1999, pp. 25–28.
[87] O. Takahashi, S. Dhong, M. Ohkubo, S. Onishi, R. Dennard, R. Hannon, S. Crowder, S. Iyer, M. Wordeman, B. Davari, W. B. Weinberger, and N. Aoki, "1 GHz fully pipelined 3.7 ns address access time 8K $\times$ 1024 embedded DRAM macro," in *Proc. Int. Solid State Circuits Conf.*, Feb. 2000, pp. 396–397.
[88] D. R. Bearden, D. G. Caffo, P. Anderson, P. Rossbach, N. Iyengar, T. A. Petersen, and J.-T. Yen, "A 780 MHz PowerPC microprocessor with integrated L2 cache," in *Proc. Int. Solid State Circuits Conf.*, Feb. 2000, pp. 90–91.
[89] F. Nemati and J. D. Plummer, "A novel thyristor-based SRAM cell (T-RAM) for high-speed, low-voltage, giga-scale memories," in *IEDM Tech. Dig.*, 1999, pp. 283–289.
[90] C. Lage, J. D. Hayden, and C. Subramanian, "Advanced SRAM technology—The race between 4T and 6T cells," in *IEDM Tech. Dig.*, 1996, pp. 271–272.
[91] K. Itoh, A. R. Fridi, A. Bellaouar, and M. I. Elmasry, "Deep sub-V, single power-supply SRAM cell with multi V_T, bosted storage node and dynamic load," in *Proc. Symp. VLSI Technology*, June 1996, pp. 132–133.
[92] A. R. Fridi, P. M. Solomon, D. J. Frank, S. Reynolds, D. Pearson, and M. I. Elmasry, "A 0.22 μm CMOS 0.65 V 500 MHz 64 Kb SRAM macro,", unpublished, 1998.
[93] T. H. Meng, B. M. Gordon, E. K. Tsern, and A. C. Hung, "Portable video-on-demand in wireless communication," *Proc. IEEE*, vol. 83, pp. 359–380, Apr. 1995.
[94] D. Cheseboro, J. Adkinson, L. Clark, S. Eslinger, M. Faucher, S. Holmes, R. Mallette, E. Nowak, E. Sengele, S. Voldman, and T. Weeks, "Overview of gate linewidth control in the manufacture of CMOS logic chips," *IBM J. Res. Dev.*, vol. 39, no. 1/2, pp. 198–200, 1995.
[95] L. Su *et al.*, "A high performance 0.08 μm CMOS," in *Proc. Symp. VLSI Technology*, June 1996, p. 12.
[96] Y. Yang *et al.*, "Sub-60 nm physical gate length SOI CMOS," in *IEDM Tech. Dig.*, 1999, pp. 431–434.
[97] J. R. Brews, "Physics of the MOS transistor," in *Applied Solid State Science*. New York: Academic, 1981, pp. 1–120.

David J. Frank (Member, IEEE) received the B.S. degree from the California Institute of Technology, Pasadena, in 1977, and the Ph.D. degree in physics from Harvard University, Cambridge, MA, in 1983.

He is currently with the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he is a Research Staff Member. He has served on technical program committees for the International Electron Devices Meeting and the Si Nanoelectronics Workshop. He has authored or coauthored over 70 technical publications and holds six U.S. Patents. His studies and recent work include nonequilibrium superconductivity, modeling and measuring III–V devices, exploring the limits of scaling of silicon technology, the modeling of innovative Si devices, analysis of CMOS scaling issues such as discrete dopant effects and short-channel effects associated with high-k gate insulators, investigating the usefulness of energy-recovering CMOS logic and reversible computing concepts, and low-power circuit design. His research interests include superconductor and semiconductor device physics, modeling and measurement, circuit design, and percolation in two-dimensional systems.

Robert H. Dennard (Fellow, IEEE) was born in Terrell, TX, in 1932. He received the B.S. and M.S. degrees in electrical engineering from Southern Methodist University, Dallas, TX, in 1954 and 1956, respectively, and the Ph.D. degree from the Carnegie Institute of Technology, Pittsburgh, PA, in 1958.

He then joined IBM Research Division, where his early experience includes the study of new devices and circuits for logic and memory applications and the development of advanced data communication techniques. Since 1963, he has been with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, where he has been involved in microelectronics research and development. His primary research interests has been in MOSFETs and integrated digital circuits that use them. In 1967, he invented the dynamic RAM memory cell used in most computers today. With others, he developed the concept of MOSFET scaling in 1972.

Dr. Dennard was appointed an IBM Fellow in 1979 and was elected to the National Academy of Engineering in 1984. He was inducted into the National Inventors Hall of Fame and became a Member of the American Philosophical Society in 1997. He received the IEEE Cledo Brunetti Award in 1982, the National Medal of Technology from President Reagan in 1988 for his invention of the one-transistor dynamic memory cell, the IRI Achievement Award from the Industrial Research Institute in 1989, and the Harvey Prize from Technion, Haifa, Israel, in 1990.

Edward Nowak (Member, IEEE) received the B.S. degree in physics from the Massachusetts Institute of Technology, Cambridge, in 1973, and the M.S. and Ph.D. degrees from the particle theory group at the University of Maryland, College Park, in 1974 and 1978, respectively.

Following post-doctoral research at New York University, he joined IBM's technology development group, Essex Junction, VT, to work on 1-Mb DRAM, and then began work on sub-half-micron MOSFETs for logic in 1984. He has contributed to numerous high-speed CMOS projects from 1.0-μm to 0.1-μm scales. He invented on 32 U.S. patents in the areas of CMOS circuits, devices and processes, and has authored numerous papers in these areas. He is currently engaged in the pursuit of sub-one-volt device designs and continues work on high-speed CMOS device design.

Paul M. Solomon (Fellow, IEEE) was born in Cape Town, South Africa. He received the B.Sc. degree in electrical engineering from the University of Cape Town, South Africa, in 1968 and the Ph.D. degree from the Technion, Haifa, Israel, in 1974.

Since 1975, he has been a Research Staff Member at the I.B.M. T. J. Watson Research Center, Yorktown Heights, NY, where his interests have been in the field of high-speed semiconductor devices. He has contributed to the physics of transport in semiconductors and has taught the physics of high-speed devices at Stanford University, Stanford, CA. He has also contributed to the theory of scaling bipolar transistors to very small dimensions and has developed methodologies to compare the performance of high-speed semiconductor devices. The design of high-speed semiconductor logic devices has been a continuing topic, ranging from self-aligned bipolar transistors through novel heterostructure field effect transistors and, more recently, to novel CMOS device concepts.

Dr. Solomon is a Member of the APS.

Yuan Taur (Fellow, IEEE) received the B.S. degree in physics from National Taiwan University, Taipei, Taiwan, R.O.C., in 1967 and the Ph.D. degree in physics from University of California, Berkeley, in 1974.

From 1975 to 1979, he was with NASA, Goddard Institute for Space Studies, NY, working on low-noise Josephson junction mixers for millimeter-wave detection. From 1979 to 1981, he was with Rockwell International Science Center, Thousand Oaks, CA, working on II–VI semiconductor devices for infrared sensor applications. Since 1981, he has been with the Silicon Technology Department of IBM Thomas J. Watson Research Center, Yorktown Heights, NY, where he was Manager of Exploratory Devices and Processes. He has served on the technical program committees and as a Panelist at the Device Research Conference, International Electron Device Meeting, and as Rump Session Chairman and Secretary at the Symposium on VLSI Technology. His recent work includes Latchup-free 1-μm CMOS, self-aligned TiSi2, 0.5-um CMOS and BiCMOS, shallow trench isolation, 0.25-μm CMOS with n^+/p^+ poly gates, SOI, low-temperature CMOS, and 0.1-μm CMOS. He has authored or coauthored over 100 technical papers, holds 10 U.S. patents, and coauthored *Fundamentals of Modern VLSI Devices* (Cambridge, U.K.: Cambridge Univ. Press, 1998).

Dr. Taur is the Editor-in-Chief of the IEEE ELECTRON DEVICE LETTERS. He received four Outstanding Technical Achievement Awards and six Invention Achievement Awards during his IBM career.

Hon-Sum Philip Wong (Senior Member, IEEE) received the B.Sc. degree from the University of Hong Kong in 1982 and the Ph.D. degree in electrical engineering from Lehigh University, Bethlehem, PA, in 1988.

He joined the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, in 1988, where he is now Senior Manager of the Exploratory Devices and Integration Technology Department. Since 1993, he has been working on device physics, fabrication, and applications of nanoscale CMOS devices. His recent work has been on the physics and fabrication technology of double-gate and back-gate MOSFETs for CMOS technologies toward the 25-nm channel length regime. In 1997, he reported the first successful fabrication of a self-aligned double-gate MOSFET using a pattern-constrained selective epitaxial growth technique. In the applications arena, his work has been on solid-state imaging. His recent work has been imaging devices using CMOS technologies. From 1988 to 1992, he worked on the design, fabrication, and characterization of a high-resolution, high-color–fidelity CCD image scanner for art work archiving. These scanners are now in use at several premier museums around the world. His research interests have been in electron device physics, device simulation and modeling, microelectronics fabrication technology, applications of microelectronic systems, and solid-state imagers.

Dr. Wong serves on the IEEE Electron Devices Society as Chair of the VLSI Circuits and Technology Committee and a Member of the Publication Committee.