Power-constrained CMOS scaling limits

by D. J. Frank

The scaling of CMOS technology has progressed rapidly for three decades, but may soon come to an end because of power-dissipation constraints. The primary problem is static power dissipation, which is caused by leakage currents arising from quantum tunneling and thermal excitations. The details of these effects, along with other scaling issues, are discussed in the context of their dependence on application. On the basis of these considerations, the limits of CMOS scaling are estimated for various application scenarios.

1. Introduction

For the past 25 years Si CMOS technology has been advancing along an exponential path of shrinking device dimensions, increasing density, increasing speed, and decreasing cost. Throughout that time people have been proposing limits to this progress, based primarily on physical phenomena, many of which have fallen by the wayside. This work describes the present state of understanding of these limits, and seeks to add to that understanding by considering the way in which application-dependent power-dissipation constraints enter into the setting of limits.

There are basically two types of power dissipation in a CMOS circuit: dynamic and static. The dynamic power is usefully expended, since it is associated with the switching of logic states that is central to performing logic operations. Dynamic power is proportional to $CV_{DD}^2 f$, where C is the capacitance, V_{DD} is the supply voltage, and f is the clock frequency. This power dissipation is in direct proportion to the rate of computation, and so can be adjusted to meet application power requirements by adjusting the computation rate. It can also be adjusted, to a more limited extent, by adjusting the supply voltage.

Static power, on the other hand, is associated with the holding or maintenance of logic states between switching events. This power is due to leakage mechanisms within the device or circuit, and so is wasted because it does not contribute to computation. Unfortunately, leakage is unavoidable, and the mechanisms are rapidly increasing in severity as scaling proceeds. By considering these mechanisms in conjunction with the power-dissipation requirements of different applications, it has been found that static power plays a central role in determining how far scaling can go, and that there is no single "end to scaling." Rather, there is a wide range of ends to scaling, corresponding to optimized technologies for different applications [1].

The organization of the paper is as follows. The next section summarizes background information on CMOS scaling and on the physical effects that limit scaling. The third section describes an analysis of static-power dissipation in CMOS circuitry and couples that analysis to application-dependent power-dissipation constraints to provide an estimate of how the limits of scaling vary with application. The fourth section discusses some of the

0018-8646/02/$5.00 © 2002 IBM

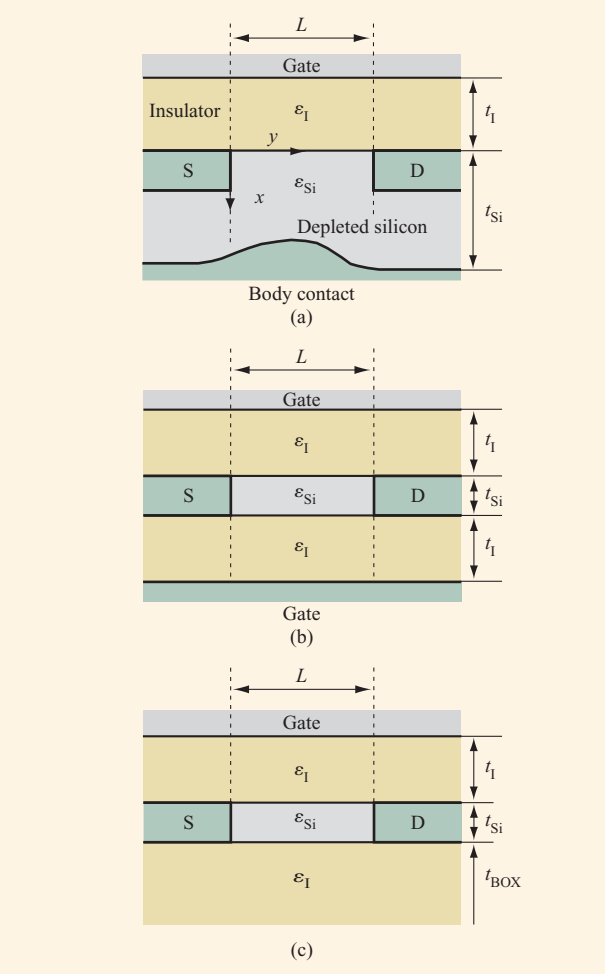

Figure 1

MOSFET cross-sectional diagrams: (a) bulk MOSFET; (b) double-gate MOSFET; (c) SOI MOSFET. The ε's are dielectric constants of the respective layers, and the t's are layer thicknesses as shown. S and D respectively indicate heavily doped source and drain regions. BOX = buried oxide.

consequences of the preceding analysis, and the final section is a conclusion.

2. Scaling issues

Device structures

As VLSI technology approaches the limits of CMOS scaling, there are three primary device structures under consideration. **Figure 1** illustrates these devices schematically, and serves to define the dimensional variables that are used here. The bulk MOSFET shown in Figure 1(a) is the conventional and most widespread FET structure. The double-gate MOSFET (DG-FET) shown in Figure 1(b) is a theoretical and exploratory device with many different experimental variations. From a theoretical point of view, it has been shown [2, 3] that this structure potentially has better short-channel effects than a bulk MOSFET of similar channel length, especially at the limits of scaling. Finally, a silicon-on-insulator (SOI) MOSFET is shown in Figure 1(c). This last device structure occupies the middle ground between the previous two cases and can display quite complex behavior; however, to avoid getting lost in the details, the present analysis adopts the simplification that SOI MOSFETs can be divided into two categories: partially depleted (PD) and fully depleted (FD) (depending on how far the doping in the thin Si channel region is depleted), and these will be lumped in with the bulk and DG-FETs, respectively.

From processing and electrostatic points of view, bulk and PD-SOI are very similar MOSFET structures. The biggest difference is the floating-body effect in PD-SOI, which occurs in devices without body contacts when majority carriers collect in the body of the FET, forward-biasing the body relative to the source and causing the effective threshold voltage to shift. This effect can be accommodated by circuit design or countered by use of a body contact, making scaling limit considerations very similar for these two cases.

Although it has been well demonstrated that thin FD-SOI devices do not scale as well as DG-FETs [3, 4], it makes some sense to consider these devices as similar because they have similar processing issues regarding the thin Si layer and the ohmic contacts, and because they have similar tunneling leakage considerations. In considering the results, however, it must be remembered that FD-SOI devices cannot be fabricated to the same dimensions as DG-FETs—the channel length must be longer or the Si thinner to achieve the same short-channel behavior.

Scaling of these device structures is a well-explored science [1, 5–7], in which the dimensions and voltages are all decreased approximately in proportion to one or more scaling parameters while the doping is increased in similar proportion, as described in the preceding references in more detail. When this scaling works, succeeding generations of technology have denser, higher-performance circuits without too much increase in power density. The limits of this scaling process are caused by various physical effects that do not scale properly, including quantum-mechanical tunneling, the discreteness of dopants, voltage-related effects such as subthreshold swing, built-in voltage and minimum logic voltage swing, and application-dependent power-dissipation limits.

One of the important goals of scaling is to maintain adequate gate control over the drain current. It has recently been shown that there is an accurate electrostatic scale length Λ_1 for the potential in the channel of an FET, such that the L/Λ_1 ratio is a good measure of the 2D

effects in the FET [8]. This Λ_1 is given implicitly as the largest solution of

$$0 = \varepsilon_{Si} \tan(\pi t_I/\Lambda_1) - \varepsilon_I \tan\pi(1 - t_{Si}/\Lambda_1) \quad (1)$$

for bulk devices and

$$0 = \varepsilon_{Si} \tan(\pi t_I/\Lambda_1) - \varepsilon_I \tan\left[\frac{\pi}{2}(1 - t_{Si}/\Lambda_1)\right] \quad (2)$$

for symmetric DG devices, using the variables defined in Figure 1. **Figure 2** shows the dependence of various FET characteristics on the L/Λ_1 ratio. On the basis of this analysis, it appears that $L/\Lambda_1 \sim 1.5$ is a good nominal design point for most FET technologies, allowing adequate room for tolerances of up to ±30%, provided

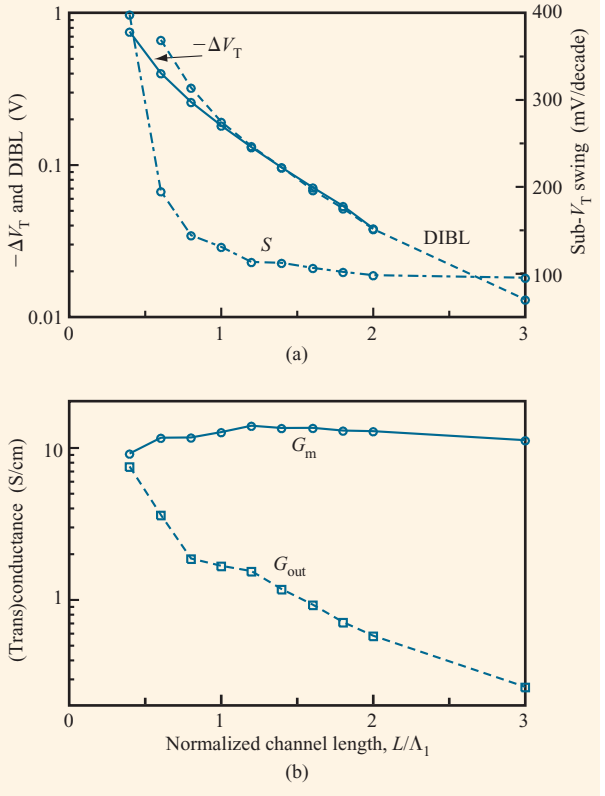

Figure 2

Dependence of 2D effects on the L/Λ_1 ratio: (a) ΔV_T, DIBL, and subthreshold swing (S) versus L/Λ_1; (b) transconductance (G_m) and output conductance (G_{out}) versus L/Λ_1. Based on 2D FIELDAY simulations of an idealized bulk MOSFET with $\Lambda_1 = 13.6$ nm ($t_{ox} = 1.5$ nm, $t_{Si} = 10$ nm). The ΔV_T is determined at $V_{DS} = 0.05$ V, the 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. From [1], reproduced with permission; ©2001 IEEE.

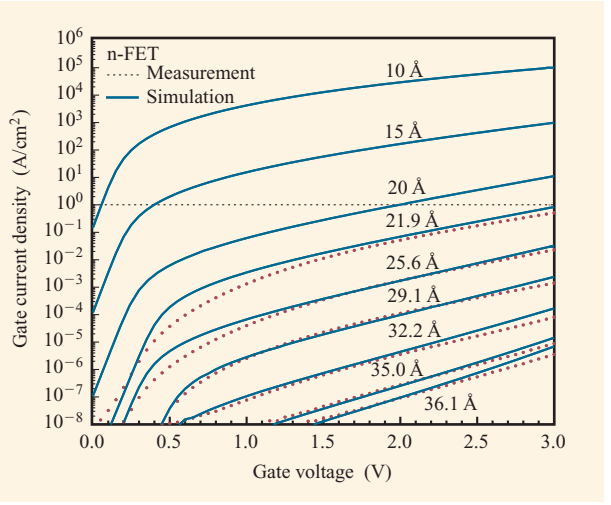

Figure 3

Calculated (curves) and measured (dots) results for tunnel currents from inversion layers through thin oxides. Adapted from Lo et al. [9]. From [7], reproduced with permission; ©1999 IEEE.

that some of the threshold voltage (V_T) roll-off is compensated by the use of halo doping. It may also be possible to improve the drain-induced barrier-lowering (DIBL) curve by use of source–drain asymmetry, such as a larger source side halo or a SiGe source contact. These techniques may shift the peak barrier in the channel closer to the source, making the subthreshold current less sensitive to drain voltage and enabling a slightly lower L/Λ_1 design point.

Tunneling effects

One of the most important effects that limit scaling is the quantum-mechanical tunneling of carriers through the energy barriers in the device. This tunneling results in leakage current, which increases power dissipation and decreases logic operating margins. There are three forms of this leakage of particular importance: tunneling through the gate insulator, band-to-band (Zener) tunneling between the body and drain, and direct source-to-drain tunneling through the channel barrier. Oxide tunneling between gate and channel is the most prominent and well known of these leakage currents, and is illustrated in **Figure 3**. In n-FETs this current is due to the tunneling of electrons from the channel to the gate. In p-FETs the tunneling current may be due to hole tunneling from channel to gate for very thin oxides (<1.5 nm) and low voltages, but at higher bias it is more often due to tunneling of electrons from the valence band of the gate into the conduction band of the body. This asymmetry

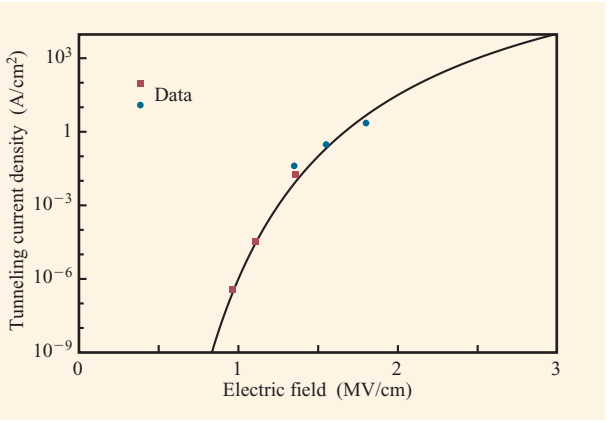

Figure 4

Band-to-band tunneling current density versus peak electric field for 1 V reverse bias. Adapted from [10]; data is from [11] and [12]. From [1], reproduced with permission; ©2001 IEEE.

exists because the valence-band barrier height is ~5 eV, while the conduction-band barrier is only ~3 eV.

There is much recent work aimed at reducing the gate tunneling problem by changing to a higher-permittivity (k) gate insulator. Currently the only successful insulators of this sort are Si oxy/nitride composites. A high-k gate insulator is characterized by three thicknesses: its physical thickness t_I, its equivalent-oxide tunneling thickness t_{oxTeq}, and its equivalent-oxide capacitive thickness t_{oxCeq}. By definition, all three are equal for SiO_2. Although t_I is larger than the equivalent SiO_2 film thickness for most high-k dielectrics, the goal is to find an insulator with the property that its t_{oxCeq} is significantly less than its t_{oxTeq} when t_{oxTeq} is equal to the minimum SiO_2 thickness. This would enable further scaling, since, at least initially, when the gate insulator permittivity varies, 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) [1]. It should be noted that gate depletion also plays a role in these considerations, since it increases the effective capacitive thickness. This tends to favor the use of high-k dielectrics in combination with metal gates (which have negligible depletion).

The second important source of tunneling leakage current is band-to-band tunneling between the body and drain of an FET. This current is strongly dependent on the electric field, as shown in **Figure 4**, which is based on reverse-current measurements in the emitter–base junctions of bipolar transistors [10–12]. Since direct band-to-band tunneling depends on conduction-band states being lined up with valence-band states, it can be avoided in undoped-channel DG-FETs if $V_T + V_{DS} \le E_G$, while in bulk MOSFETs the equivalent condition is $V_{DS} - V_{BS} \le 0$, where V_{DS} is the drain-to-source voltage, V_{BS} is the body-to-source voltage, and E_G is the bandgap. Thus, the condition can readily be avoided in DG-FETs at low voltage, but for bulk FETs, it requires forward body bias exceeding the supply voltage, V_{DD}. At low temperature the latter might be an interesting option [1], but it is unlikely that it would be applied to anything except very-high-performance computing. Indirect band-to-band tunneling through deep traps in the depletion region often dominates over direct tunneling, and can readily violate the preceding voltage condition. To reach the limits of scaling, it will probably be necessary to find ways to eliminate such traps.

An analytic approximation for band-to-band tunneling current in a 1D geometry may be obtained by assuming that the tunneling current varies locally as $J(x) \sim e^{-B/F_{eff}(x)}$, where $F_{eff}(x) = E_G/(x_2 - x)$, x is the starting point for tunneling, x_2 is the point at which the particle would reappear in the opposite band on the other side of the junction, and B is a fitting parameter; and then approximating the integral over those xs for which tunneling is possible. For an abrupt one-sided junction, this yields

$$J_{B2B}(V_{DB}, F_{max}) \approx \frac{1.4 \times 10^{10}\ \mathrm{A/cm^2}}{F_{max}} \sqrt{E_G(E_G + V_{DB})}\, e^{-\alpha}\left[\frac{e^{bu/(u+1)} - 1}{b}\right], \quad (3)$$

where $b = 2.9 + 1.14\alpha$, $u = \sqrt{V_{DB}/E_G}$, $\alpha = (56.3/F_{max})\sqrt{1 + V_{DB}/E_G}$, V_{DB} is the drain-to-body voltage, and F_{max} is the maximum field in units of MV/cm at the junction edge. The numerical parameters are calibrated to the data in Figure 4. Although the initial assumption is fairly crude, this functional form has reasonable voltage dependence, and it is used in the calculation in Section 3.

The final tunneling current of possible concern is direct source-to-drain tunneling, through the channel barrier. This contribution can become observable for channel lengths shorter than 20 nm, especially at low temperature [13], but most recent analyses show that it becomes problematic at room temperature only for channel lengths below ~10 nm [14]. Since FETs can achieve such short channel length only for very-high-performance, high-power-density applications, this extra tunneling current turns out to be comparatively negligible for the cases of interest.

Discrete doping effects

Another physical effect that may limit scaling is the discreteness of the dopant atoms. Although the average concentration of doping is quite well controlled by the standard ion implantation and annealing processes, these

processes do not control exactly where each dopant ends up. Consequently there is randomness at the atomic scale, resulting in spatial fluctuations in the local doping concentration, and these in turn cause device-to-device variation in MOSFET threshold voltages. As MOSFET technology nears the end of scaling, it will be readily possible to make devices with fewer than 100 dopant atoms controlling 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, threshold variation may very well become quite large, making the design of robust circuits very difficult.

Many workers have investigated the effects of these doping fluctuations on the V_T of MOSFETs, the most quantitatively accurate of which use stochastically placed dopants in full 3D MOSFET simulations to fully resolve the effects of dopant placement [15–18]. **Figure 5** shows an example of the statistical variation expected to occur in an 11-nm bulk MOSFET due to random dopant placement. These particular 3D simulations were carried out using the FIELDAY program [19] coupled with a preprocessor [17] to randomly place the dopants. They represent the worst-case (20% short) result for a nominal 14-nm design point which was scaled from the published 25-nm design of Taur et al. [10]. It seems clear that such a design point will be unusable from a circuit point of view because of the very wide variation in threshold voltage.

However, it is difficult to predict the extent to which this effect will limit scaling, since there are several approaches to reducing the effect, and more may be discovered. For bulk devices, the most obvious approach is to move the dopants in the body back away from the surface using highly retrograde channel doping profiles. Stochastic simulations confirm that such profiles can yield significantly (more than two times) lower V_T uncertainty than uniformly doped channels [17, 18]. This is because the doping fluctuations are moved farther 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. The best way to eliminate these fluctuations is to remove the doping, and this may be possible in DG-FETs, if the threshold can be set by the gate work function instead of by doping. Even if they do require doping, they may not require very much doping to obtain the desired threshold, and so the fluctuations may also be lower [20].

Voltage effects

There are several voltage-related issues that affect the scaling of CMOS, the most important of which is that V_T cannot be fully scaled because the off-current I_{off} of the FET is constrained by application considerations. I_{off} is related to V_T by

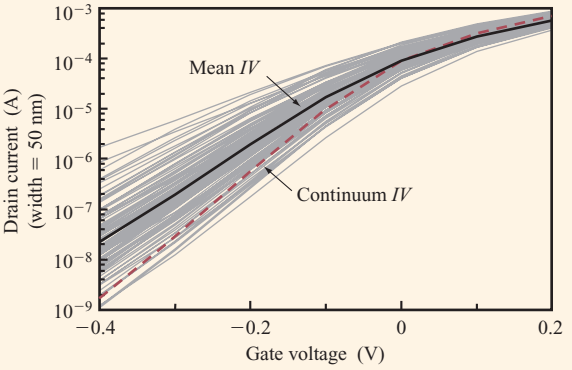

Figure 5

Simulated IV curves for 100 different 11-nm-channel-length bulk MOSFETs with discretely placed dopants. Each gray curve corresponds to a different random placement of the dopants in keeping with the designed average doping profiles. The solid black curve is the geometric average of the curves, while the dashed curve is the expected IV curve based on continuum doping profiles.

$$I_{off} \cong I_{VT} 10^{-V_T/S}, \tag{4}$$

where S is the subthreshold swing and I_{VT} is the current at which V_T is defined. Since $S \cong (\ln 10)\eta kT/e$, where η is the ideality, k is Boltzmann's constant, and T is the temperature, the only way to scale 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 low-to-moderate-power applications, I_{off} may be in the 10^{-7}-to-10^{-4}-A/cm range, resulting in minimum V_Ts between 0.54 and 0.27 V, respectively, assuming I_{VT} = 0.1 A/cm and S = 90 mV/decade. (These are worst-case thresholds; nominal threshold voltages must be set higher to allow for manufacturing tolerances.) Since DG-FETs generally have smaller subthreshold swing, perhaps 70 mV/decade at room temperature, their thresholds and hence the supply voltages can be scaled further.

Two application considerations constrain I_{off}: It cannot be so high that the circuit does not function, and the total power dissipation associated with the leakage must be tolerable for the given application. The latter constraint is usually more important because of the enormous device density on modern chips. The most effective way of dealing with this constraint is by optimizing the V_T and V_{DD} for the desired speed and power dissipation. This optimization has been well studied, especially in the low-power regime [21–23], where the effects of process and supply variations are quite important. The results of one such study [23] are shown in **Figure 6**, which illustrates the dependence of the optimum design points on activity

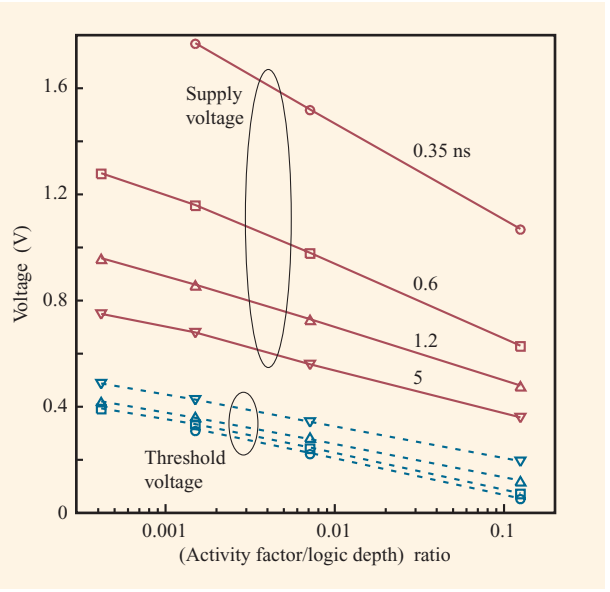

Figure 6

Optimum supply voltage and threshold voltage versus activity factor-to-logic depth ratio for four different delay constraints, based on simulations of 0.1-μm CMOS technology. Threshold voltage here is defined as the gate voltage at which $\sqrt{I_{D,on}}$ extrapolates to zero. Data is from [23]. From [1], reproduced with permission; ©2001 IEEE.

factor and logic depth. These particular optimizations are for 0.1-μm static CMOS arithmetic circuits with realistic tolerances, but the optimal voltages should not vary too much as technology is scaled (provided the delay target is suitably scaled). Each point in the figure represents an independent optimization of both the supply voltage and the threshold voltage. As shown, the optimum voltages depend strongly on activity factor and logic depth, so that a wide range of V_Ts and V_{DD}s are needed to satisfy the requirements of a range of applications. Note that these supply voltages are much larger than the theoretical minimum supply voltages of ~3–4 kT required for self-consistent logic [1].

A secondary voltage-scaling issue is the bandgap E_G, which does not scale since it is a property of the semiconductor. Although the nonscaling of the bandgap complicates device design by increasing junction fields and depletion depths, it does not truly limit device scaling, since its effects can be countered by higher doping or even forward biasing of the body.

3. Power-constrained scaling limits

Although most of the nonscaling effects that have been described have the potential of halting CMOS scaling at the point at which they cause circuits to cease functioning, that is not the important scaling limit. As mentioned in connection with I_{off}, the most significant scaling limit is created by the power dissipation associated with the various leakage mechanisms. This limit depends on application, since different applications can tolerate different amounts of static leakage power, so that there is no single end to scaling, but rather there are different optimum ends to scaling for different applications. High-power, high-performance servers can accept much higher static leakage dissipation than portable battery-powered devices, and so the former can be more aggressively scaled than the latter.

To better illuminate this point, the approximate scaling limits for various application classes have been calculated, in the same manner as Reference [1], and this data is presented in **Table 1**, for both bulk-like MOSFETs and DG-FETs. This table is intended to show the general trends and dependencies of these limits, rather than exact values. Total power density is the overriding parameter, and the leakage mechanisms are each allocated a fraction of the total power. More detailed optimizations are needed to more precisely determine these fractions, but although such optimizations are likely to change the fractions somewhat, the results in the table are only logarithmically dependent on these values, so the final conclusions should not change very much.

There are two types of circuit application in this table: SRAM cells, for which it is assumed that essentially all of the power is static (i.e., very little activity), and logic circuits, for which it is assumed that the switching activity is at least a few percent, and the static power is about a third of the total power. The latter case implicitly assumes 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. The peripheral circuitry of an SRAM is thought of as being included in the active logic category. As is the current practice, it is expected that multiple technologies may be present on the same chip; the table is thought of as addressing the requirements for the dominant technology on a given section of a chip. See Section 4 for more discussion of this issue.

The methodology used to create the table is described in detail in [1], but the essence is as follows. Starting with an approximate channel length, the fraction of the power allocated to subthreshold dissipation is used to determine V_T. The fraction allocated to gate current is used to calculate the insulator thickness t_I (an oxy-nitride gate stack is assumed here). The fraction allocated to band-to-band tunneling (together with the V_T in the case of DG-FETs) is used to calculate the t_{Si}. Given t_I and t_{Si}, the

Table 1 Estimated scaling limits for MOSFET design parameters as a function of application class and device structure. 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 V_{DD} being independent variables.

Device type	*Application*	*T* (°C)	*Power* (W/cm^2)	V_{DD} (V)	I_{off} (nA/μm)	V_{Tn} (mV)	t_{oxTeq} (nm)	t_{Si} (nm)	L_{nom} (nm)
Bulk	High performance	85	1000	0.8–1.2	3100–2600	102	0.9–1.0	6–8.5	13–17
		85	100	0.8–1.0	370–340	185	1.1–1.2	8–9	16–18
Bulk	Medium–high performance	85	10	0.6–1.0	50–40	270	1.2–1.4	8–11	16–21
Bulk	Moderate performance	85	1.0	0.6–1.0	6–4.5	360	1.4–1.6	9–12	19–24
Bulk	Low power	65	0.05	0.7–0.9	0.32–0.28	450	1.7–1.8	11–13	24–27
Bulk	Ultralow power	40	<0.001	0.7–1.0	<0.0075	550–710	2.1–2.6	13–19	28–39
Bulk	Moderate-performance SRAM	85	5–1	0.9–1.2	60–10	260–310	1.3–1.6	10–13	20–26
	Low-power SRAM	65	0.1–0.01	0.9–1.2	1.5–0.15	380–470	1.6–2.0	12–16	25–32
	Ultralow-power SRAM	40	0.0001	1.2	0.0018	590	2.4	20	39
DG-FET	High performance	85	10000	0.8	28000	37	0.76	4	12
		85	1000	0.8–1.2	3100–2200	110–125	1.0–1.1	4	13–14
		85	100	0.8–1.0	340–280	195	1.2	4	15
DG-FET	Medium–high performance	85	10	0.6–1.0	50–30	270	1.3–1.4	4	16
DG-FET	Moderate performance	85	1.0	0.6–1.0	5–4	340	1.5–1.6	4–6	17–22
DG-FET	Low power	65	0.05	0.7–0.9	0.25	420	1.8–1.9	4–7	19–24
DG-FET	Ultralow power	40	<0.001	0.7	<0.006	510–630	2.1–2.5	4–9	21–32
		40	<0.001	1.0	<0.007	490–620	2.2–2.6	13–19	36–49
DG-FET	Moderate-performance SRAM	85	5–1	0.9–1.2	50–10	260–290	1.4–1.6	4–9	16–26
	Low-power SRAM	65	0.1–0.01	0.9–1.2	1.2–0.2	370–410	1.7–2.0	5–14	20–38
	Ultralow-power SRAM	40	0.0001	1.2	0.002	510	2.4	20	49

scale length Λ_1 is computed, from which a more accurate estimate of the nominal channel length is determined. This procedure is iterated until converged. There are a few differences from [1]: 1) An oxy-nitride insulator ($\varepsilon = 6$) is assumed rather than Al_2O_3; 2) SRAM cells are treated as 100% static power dissipation (60% subthreshold, 30% gate current, and 10% band-to-band), 3) Equation (3) is used for the band-to-band tunneling to account more accurately for the bias voltages; 4) the DG scaling of drain electric field relative to a simulated 14-nm device has been improved; and 5) the source-to-drain tunneling limits and maximum I_{off} constraints have been removed. In spite of these improvements, the results are not very different from those in [1].

The table clearly reveals the dependence of scaling limits on application power requirements. As one moves from high-power to low-power applications, the shrinking leakage requirements cause the minimum allowed nominal channel length for bulk MOSFETs to increase three times, from ~13 nm to ~39 nm, while t_{oxTeq} increases from 0.9 nm to 2.6 nm, corresponding to tunneling current densities from 15 kA/cm^2 to 70 μA/cm^2 (at 1 V), respectively. The channel lengths of the DG-FETs increase from 12 nm to 49 nm, and show up to a 30% scaling advantage over bulk, with the largest advantage occurring for intermediate power levels and low V_{DD}, where it is equivalent to an entire generation of scaling. The advantage is lost for the high-V_T, high-V_{DD} cases, where the DG-FET is more affected by body-to-drain tunneling, though this may be an artifact of more abrupt doping profiles in the DG-FET. The advantage is also reduced for the high-power DG devices, partly because the oxide tunneling power constraint necessitates slightly thicker (~0.1-nm) gate insulators (because there is twice as much gate area per cm^2 of Si, for the assumed geometry) and partly because of the 4-nm minimum Si thickness that was imposed for the sake of tolerance control. To be fair, though, discrete doping issues may very well prohibit the bulk designs below 20 nm, which greatly increases the advantage of the DG-FETs. On the other hand, the DG-FET design points require halo-like V_T roll-off compensation and metal gates with suitable work functions to set V_T, neither of which are known processes, making the DG designs more speculative than the bulk MOSFET designs, which are better understood. A further consideration is that if a FinFET [24] geometry (in which DG-FETs are built on the sidewalls of vertical Si "fins") were assumed for the DG-FET, rather than a planar geometry, the oxide area

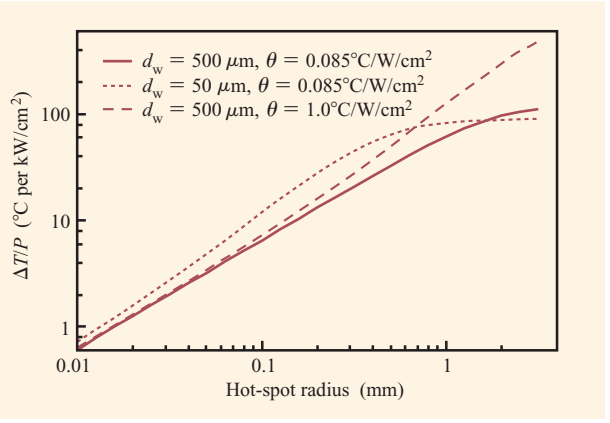

Figure 7

Coefficient of peak temperature rise of a region with higher power circuitry as a function of the radius of the region, for three different heat-sink assumptions. This is the relative temperature rise divided by the relative increase in power density compared to the rest of the chip. d_w = silicon wafer thickness; θ = thermal contact resistance.

per cm^2 of Si could be even more than twice that of bulk CMOS, requiring a still thicker gate oxide to hold tunneling dissipation in check. Since DG-FET gate capacitance per cm^2 of Si may also be at least twice that of bulk, the constraint on dynamic-power density forces the use of lower-clock-frequency, narrower devices (with their attendant tighter logic-gate pitch, yielding shorter interconnects and lower wiring capacitance) and, if margin considerations permit it, lower supply voltages.

4. Discussion

Many aspects of the preceding analysis deserve comment, but for the sake of brevity only a few are touched on here, including the question of whether the power targets are achievable, considerations involved in mixing higher-performance logic into lower-performance chips, and some comments about the uncertainties of the calculations. For discussion of various other issues, see [1].

Implicit in the static-power allocations of the table is the assumption that the active power can always be adjusted to be 60–70% of the total power-density constraint. At the very highest power density (10 kW/cm^2), this requires very active, heavily loaded circuits, such as clock drivers, data-bus drivers, or off-chip I/O drivers. Random logic does not usually reach this power level. Consequently, if the most scaled FETs are used for logic, the power-dissipation proportions will probably be different, perhaps something like 3000 W/cm^2 static power and 500 W/cm^2 dynamic, for a total of 3500 W/cm^2.

On the other hand, moving down the power scale to less aggressive technology, it should be relatively easy for even low-power technology to reach active-power densities of ~100 W/cm^2. Consequently, at the low-power end the challenge is to get the active power down to the required levels. This is primarily a matter of circuit and system design, and several approaches have been suggested [1]:

1. Since chips almost never use all of their circuitry extremely actively, it is possible to average over the less active areas and over large areas of lower-dissipation SRAM or DRAM, thus reducing the power density as much as an order of magnitude.
2. The clock frequency can be lowered until the throughput requirements are only just satisfied, and this may enable a further reduction in V_{DD}, although V_{DD} cannot be too close to V_T because threshold variations cause too much timing uncertainty.
3. The chip can be run in bursts of power-optimized activity and turned off between bursts.
4. The chip can be designed as many special-purpose macros, each power- or energy-optimized for its specific task. The work would be shuffled among the macros, minimizing the energy consumed and increasing the averaging used in the first approach.

As was noted before, it is expected that most chips will be designed to use a mixture of technologies to meet the varying needs of the system. High-V_T devices will be used for the low-activity SRAM cells, while more highly scaled low-V_T devices will be used in critical logic paths. To keep the power usage balanced, it appears that the fraction of high-power logic devices ought to vary roughly as $\sim P_{typ}/P_{high}$, where P_{typ} is the power density of the dominant device technology and P_{high} is the power density of the high-power logic devices. This results in a total system power varying very roughly as $P_{typ}\ \ln(P_{max}/P_{typ})$, where P_{max} is the power density of the highest-power technology used. For example, one might imagine a high-performance processor in which 70% of the area is 10-W/cm^2 SRAM cells, 20% is 30-W/cm^2 logic technology, 7% is 100-W/cm^2, 2% is 300-W/cm^2, 0.7% is 1000-W/cm^2, and 0.3% is 3000-W/cm^2. Such a processor would dissipate 42 W/cm^2, which can be cooled using reasonable technology, but raises the economic question that will probably dominate the end of scaling. Does that last 1% of superhigh-performance devices on the chip contribute enough additional speed or function to the chip to justify the processing cost of adding it?

The preceding example also raises the issue of "hot spots." If the high-performance devices are concentrated together in a cluster (as well they might be), will that spot become too hot, even if the total power budget is satisfied? To address this issue, numerical solutions of the heat-flow equation have been carried out in cylindrical

geometry. **Figure 7** shows how the maximum temperature rise varies with the spot size and power density for bulk technology. Three different cases are shown, two with a very aggressive heat-sink design (water forced through etched Si fins on the back of the wafer, as in Tuckerman and Pease [25], for a thermal contact resistance θ of 0.085°C/(W/cm^2), and one with a more conventional thermal resistance of 1.0°C/(W/cm^2). The curves illustrate the importance of the high thermal conductivity of the lightly doped Si substrate (1.5 W/°Ccm^2), which overcomes the difference in heat-sink resistance for the thick-Si case. The thin-Si case actually has a higher temperature rise over most of the curve because heat cannot spread as well in the thin layer, making it more dependent on the local heat-sink properties.

Assuming that a 10°C temperature rise is about the maximum desirable for a hot spot, since it is in addition to the average temperature rise of the entire chip, it appears that a 100-W/cm^2 spot can have a diameter of about 1 mm, and a 1-kW/cm^2 spot can have a diameter of about 100 μm. These clusters could contain up to about 4×10^6 and 4×10^4 logic gates, respectively, making the former suitable for substantial computation, but the latter suitable only for small macros or a few critical paths here and there.

The accuracy of the scaling limit projections in Table 1 rests mostly on the leakage-current mechanisms discussed earlier. The 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. The t_{oxTeq} requirements are based on oxide tunneling curves that have been well measured in recent years. The final parameter needed to determine the minimum scaling dimension is the depletion depth (for bulk MOSFETs) or the Si thickness (for DG-FETs). In the present model these are determined from the band-to-band tunneling model [Equation (3)] based on the data in Figure 4. There is relatively little data here, and much sensitivity to mid-gap traps and the detailed doping profile. This area deserves much further investigation because it may play a prominent role in the end of scaling. Nevertheless, the results are not as uncertain as it may seem. Even if band-to-band tunneling were entirely removed as a mechanism, one would still end up with essentially the same limits to scaling if a realistic ideality factor η were used to set the ratio between oxide thickness and depletion depth.

5. Conclusion

The continued scaling of CMOS technology is imperiled by a variety of nonscaling physical effects, including the dependence of subthreshold behavior on temperature, quantum tunneling of carriers through the gate insulator and through the body-to-drain junction, and discrete doping effects. Several of these effects have the ability to halt the scaling of CMOS by making circuits nonfunctional, but this is not the primary limit to scaling. Rather, the most important limit is the power dissipated in the various leakage mechanisms. This leakage dissipation creates a whole range of application-dependent limits to scaling, since each application has its own constraints on the amount of leakage dissipation tolerable. This range of limits spans at least a factor of at least 3 in minimum FET dimensions and oxide thickness, creating the need for a wide range of technology at the end of scaling. Bulk and double-gate MOSFET structures have been compared at these scaling limits, and it appears at present that DG-FETs will hold an advantage in the end for most applications, but the size of this advantage depends on details of band-to-band tunneling and discrete doping effects that have yet to be thoroughly explored.

Acknowledgment

This work has benefited greatly from many useful discussions with co-workers and colleagues, including especially the collaborators on Reference [1]: Bob Dennard, Ed Nowak, Paul Solomon, Yuan Taur, and H.-S. Philip Wong.

References

1. D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, "Device Scaling Limits of Si MOSFETs and Their Application Dependencies," *Proc. IEEE* **89,** 259–288 (2001).
2. D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo Simulation of a 30 nm Dual-Gate MOSFET: How Far Can Si Go?," *IEDM Tech. Digest*, p. 553 (1992).
3. 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. Digest*, pp. 407–410 (1998).
4. R. Yan, A. Ourmazd, and K. F. Lee, "Scaling the Si MOSFET: From Bulk to SOI to Bulk," *IEEE Trans. Electron Devices* **39,** 1704–1710 (1992).
5. R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits* **SC-9,** 256–268 (1974).
6. B. Davari, R. H. Dennard, and G. G. Shahidi, "CMOS Scaling, the Next Ten Years," *Proc. IEEE* **89,** 595–606 (1995).
7. H.-S. P. Wong, D. J. Frank, P. M. Solomon, H.-J. Wann, and J. Welser, "Nanoscale CMOS," *Proc. IEEE* **87,** 537–570 (1999).
8. D. J. Frank, Y. Taur, and H.-S. P. Wong, "Generalized Scale Length for Two-Dimensional Effects in MOSFET's," *IEEE Electron Device Lett.* **19,** 385–387 (1998).
9. 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 nMOSFET's," *IEEE Electron Device Lett.* **18,** 209 (1997).
10. Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS Design Considerations," *IEDM Tech. Digest*, pp. 789–792 (1998).

11. R. B. Fair and H. W. Wivell, "Zener and Avalanche Breakdown in As-Implanted Low-Voltage Silicon N-P Junctions," *IEEE Trans. Electron Devices* **ED-23,** 512 (1976).
12. J. M. C. Stork and R. D. Isaac, "Tunneling in Base–Emitter Junctions," *IEEE Trans. Electron Devices* **ED-30,** 1527 (1983).
13. H. Kawaura, T. Sakamoto, and T. Baba, "Direct Source–Drain Tunneling Current in Subthreshold Region of Sub-10-Gate EJ-MOSFETs," *Si Nanoelectronics Workshop Abstracts*, 1999, pp. 26–27.
14. Y. Naveh and K. K. Likharev, "Modeling of 10-nm-Scale Ballistic MOSFETs," *IEEE Electron Device Lett.* **21,** 242–244 (2000).
15. H.-S. Wong and Y. Taur, "Three-Dimensional 'Atomistic' Simulation of Discrete Microscopic Random Dopant Distributions Effects in Sub-0.1 μm MOSFETs," *IEDM Tech. Digest*, pp. 705–708 (1993).
16. H.-S. P. Wong, Y. Taur, and D. Frank, "Discrete Random Dopant Distribution Effects in Nanometer-Scale MOSFETs," *Microelectron. Reliability* **38,** 1447–1456 (1998).
17. D. J. Frank, Y. Taur, M. Ieong, and H.-S. P. Wong, "Monte Carlo Modeling of Threshold Variation Due to Dopant Fluctuations," *Symposium on VLSI Technology, Digest of Technical Papers*, 1999, pp. 169–170.
18. A. Asenov and S. Saini, "Random Dopant Fluctuation Resistant Decanano MOSFET Architectures," *Si Nanoelectronics Workshop Abstracts*, 1999, pp. 84–85.
19. E. Buturla, J. Johnson, S. Furkay, and P. Cottrell, "A New 3-D Device Simulation Formulation," *NASCODE VI: Sixth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits*, Boole Press, Dublin, 1989, p. 291.
20. D. J. Frank and H.-S. P. Wong, "Simulation of Stochastic Doping Effects in Si MOSFETs," *Proceedings of the International Workshop on Computational Electronics*, 2000, pp. 2–3.
21. D. Liu and C. Svensson, "Trading Speed for Low Power by Choice of Supply and Threshold Voltages," *IEEE J. Solid-State Circuits* **28,** 10 (1993).
22. Z. Chen, J. Burr, J. Shott, and J. D. Plummer, "Optimization of Quarter Micron MOSFETs for Low Voltage/Low Power Applications," *IEDM Tech. Digest*, pp. 63–65 (1995).
23. D. J. Frank, P. Solomon, S. Reynolds, and J. Shin, "Supply and Threshold Voltage Optimization for Low Power Design," *Proceedings of the International Symposium on Low Power Electronics and Design*, 1997, pp. 317–322.
24. Y.-K. Choi, N. Lindert, P. Xuan, S. Tang, D. Ha, E. Anderson, T.-J. King, J. Boker, and C. Hu, "Sub-20nm CMOS FinFET Technologies," *IEDM Tech. Digest*, p. 421 (2001).
25. D. B. Tuckerman and R. F. W. Pease, "High-Performance Heat Sinking for VLSI," *IEEE Electron Device Lett.* **2,** 126–129 (1981).

Received October 18, 2001; accepted for publication December 19, 2001

David J. Frank *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (djf@us.ibm.com).* Dr. Frank received a B.S. degree from the California Institute of Technology, Pasadena, in 1977 and a Ph.D. degree in physics from Harvard University in 1983. Since graduation he has worked at the IBM Thomas J. Watson Research Center, where he is a Research Staff Member. His studies have included non-equilibrium superconductivity, modeling and measuring III–V devices, and exploring the limits of scaling of silicon technology. His recent work includes 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. Dr. Frank is a member of the IEEE; he has served on technical program committees for the International Electron Devices Meeting and the Si Nanoelectronics Workshop. He has authored or co-authored more than 70 technical publications and holds six U.S. patents.