Maintaining the benefits of CMOS scaling when scaling bogs down

by E. J. Nowak

A survey of industry trends from the last two decades of scaling for CMOS logic is examined in an attempt to extrapolate practical directions for CMOS technology as lithography progresses toward the point at which CMOS is limited by the size of the silicon atom itself. Some possible directions for various specialized applications in CMOS logic are explored, and it is further conjectured that double-gate MOSFETs will prove to be the dominant device architecture for this last era of CMOS scaling.

Introduction

Despite many barriers to the scaling of CMOS technology that have emerged, the exponential growth of the semiconductor industry has not only proceeded successfully for more than twenty years, but has recently actually accelerated its pace. Although gate-oxide thickness has regularly been (and continues to be) cited as an "absolute" barrier to progress, this barrier is still being defied. Various lithographic barriers based on the wavelength of visible light have fallen over the decades, and at this time state-of-the-art manufacturing facilities have already embraced lithographic exposure tools with wavelengths of 193 nm, well into the ultraviolet region of the spectrum. Other proposed barriers, such as doping, number fluctuations, and FET series resistance scaling limitations, have been discussed in the literature, but have been avoided by innovations such as ultrathin-body silicon-on-insulator (SOI) raised source/drain processes and low-barrier-height silicides. Hence, ordinary reasoning would suggest that almost any limit cited should be examined circumspectly, as new materials, clever engineering solutions, and new design methods have relentlessly broken through such barriers thus far.

However, the size of a silicon atom (or other relevant atoms) is an indisputable barrier, because any solution that does not require in the future structures of size at least comparable to the atomic scale must truly be revolutionary. Thus, the *International Technology Roadmap for Semiconductors* (ITRS) [1], which attempts to chart the scaling future for CMOS technology, must inevitably slow and finally halt, at least in the traditional sense, as the lithography scale approaches a few times atomic dimensions, or perhaps a "5-nm node." At this technology node, minimum features have dimensions of the order of 2 to 3 nm, and structures much smaller than this scale would likely be intrinsically subject to unacceptable variations. Even if extremely clever techniques should be identified to control such variations, a final barrier still presents itself at the very size of the atom (or molecule), not much beyond this scale, possibly adding a decade to

0018-8646/02/$5.00 © 2002 IBM

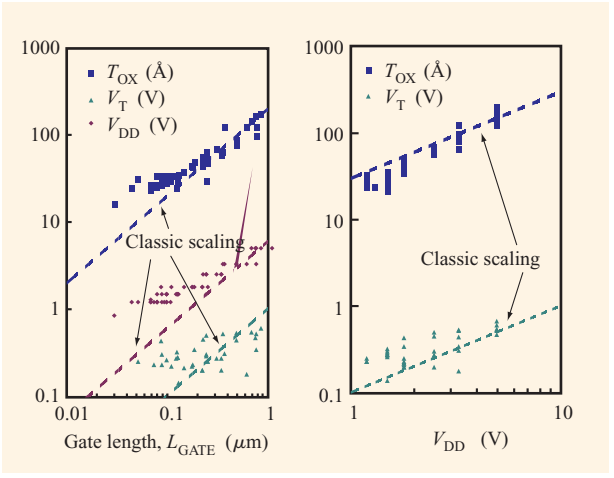

Figure 1

Published industry trends (data points) are compared to "classic" scaling (dashed curves). V_T and V_{DD} show clear signs of deviation from classic scaling with respect to L_{GATE}; the same V_T and T_{OX} data are seen to be nearly proportional to V_{DD}.

scaling assuming that the current exponential rate is preserved.

This paper reviews two decades of scaling in CMOS ULSI technology, touching very briefly on key issues that have formed worldwide concerns about continuing the process. The discussion examines some possible scenarios of "slowed" scaling as the atomic limit is approached, with particular attention to some slightly nontraditional directions that are likely to emerge as more-traditional means of leverage become more difficult to execute. Finally, the paper suggests directions in which further progress in CMOS technology may be made at the end of scaling, with particular attention to product- and manufacturing-driven issues.

Two decades of CMOS scaling

Scaling of CMOS technology has progressed relentlessly from a linewidth of 1 μm to the current 100-nm linewidth [2, 3]. Two key features characterize this era: 1) Slavish devotion to scaling by constant improvements in lithography, as described by Dennard et al. [4], and 2) a minimal rate of introduction of substantially new materials and structures. Each of these aspects is briefly explored.

While the "classic" scaling described in [4] has not been strictly followed, it has served as an essential blueprint describing the major features observed over the period from roughly 1981 to 2001. **Figure 1** shows a collection of published industry results for electrical-equivalent transistor gate-oxide thickness, T_{OX}, threshold voltage, V_T, and power-supply voltage, V_{DD}, all against reported gate length, L_{GATE}. Dashed curves show the classic scaling trajectories for these parameters as well. Two parameters of interest to this discussion which are less obvious are the drive current per unit MOSFET width, I_{DSAT}, defined as the drain current of a (unit-width) MOSFET when the gate-to-source and drain-to-source voltages are both equal to the nominal power-supply voltage, V_{DD}, and the gate capacitance per unit MOSFET width, C_{GATE}, defined as the total capacitance (per unit width) of the gate, with the source and drain grounded and the gate voltage equal to V_{DD}, and calculated from T_{OX} and L_{GATE} using

$$C_{GATE} = (\varepsilon_{OX}\varepsilon_0/T_{OX})L_{GATE} + 0.26\ \text{fF}/\mu\text{m}, \quad (1)$$

where ε_{OX} and ε_0 are respectively the relative dielectric constant of silicon dioxide ($\varepsilon_{OX} = 3.9$) and the electric permittivity of free space. Values for C_{GATE} are typically in the range of 1.0 to 1.5 fF/μm. The last term of Equation (1) takes account of capacitance from the edges of the gate electrode to the source and drain. Taking gate length as a measure of the lithography scale, one can immediately see that V_{DD}, V_T, and, to a lesser extent, T_{OX} have decreased more slowly than L_{GATE}, while I_{DSAT} has actually increased rather than remaining fixed (as in classic scaling). The right-hand side of the figure shows the same V_T and T_{OX} data as the left-hand side, except with V_{DD} as the abscissa; note that T_{OX} and V_T fall relatively close to scaling in proportion to V_{DD} (as they would in classic scaling). This suggests that the deviations from classic scaling have been driven primarily by V_{DD}, which has itself decreased more slowly than L_{GATE}. In the early part of this time span (1 μm to 0.5 μm), a reluctance to leave the widely accepted industry-standard $V_{DD} = 5.0$ V, inherited from transistor–transistor logic (TTL), substantially retarded V_{DD} reduction. As the transition to a 3.3-V standard gained momentum, an increased emphasis on performance and power resulted in circuit-board designs with a good deal of flexibility for V_{DD}; these, in turn, allowed CMOS process-technology developers the freedom to optimize V_{DD} scaling for power and performance to a greater degree. A given technology point defined by specific values of T_{OX} and L_{GATE} will nearly always deliver greater performance as V_{DD} is increased (roughly in direct proportion to V_{DD}), so as gate dielectric learning in the industry accelerated, the acceptable ratio of V_{DD}/T_{OX} increased steadily in this next era, giving rise to a continued mismatch in L_{GATE} and T_{OX} reduction rates. Thus, V_{DD} continued to decrease more slowly than L_{GATE}. The other item of note in Figure 1 is the behavior of V_T. A large scatter in V_T is seen, due in part to variability in reporting practices (nominal vs. fast-process, V_T definition, etc.) and, to a good approximation, V_T scaled in proportion to V_{DD}; this is probably largely a consequence of practical CMOS device and circuit considerations, including circuit stability, noise immunity, and engineering

of short-channel effects to acceptable levels of control. These observed behaviors are seen to give rise to a number of practical problems that pose challenges to further CMOS scaling, and these are pursued later in this discussion.

The second feature characterizing CMOS scaling over the past two decades, a measured rate of introduction of new materials, is illustrated in **Figure 2**. From 1980 to 1995, substantially new materials were introduced at the rate of about one every two or three generations. Many other, incremental changes can be found in many generations, but major new changes in materials are very difficult and costly. Substantial effort is required to introduce new materials, and great effort is required to ensure that both manufacturable and reliable integration have been attained. It is instructive to note that an accelerated rate of introduction of new materials may be suggested, as interconnects have pushed to copper and low-k dielectrics in the same time frame. This could signal an indication that the industry is approaching some "pinch points" in the continuation of scaling at the current rate of aggressiveness. The significant efforts currently under way to identify a replacement for silicon dioxide as the gate dielectric for MOSFETs and, recently, announcements regarding the introduction of silicon–germanium in CMOS technology, give further evidence of forces for change.

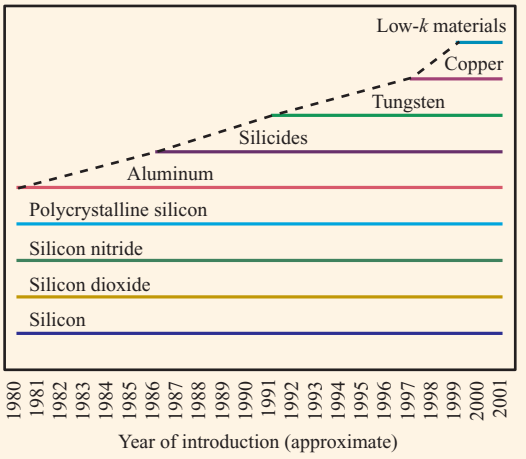

Figure 2

Approximate year of introduction of significantly new materials into mainstream CMOS technology. Note the relatively conservative rate of introduction, with just a suggestion of a recent upturn. As more "limits" of CMOS are encountered, there will be an increased impetus for an accelerated rate of introduction of new features and materials.

Approaching the atomic limit

At present, 193-nm lithography steppers are in general use. The active pursuit of advanced lithographic techniques, such as extreme ultraviolet (EUV) lithography, which makes use of light at a wavelength of 13 nm, illustrates the relentless ardor with which scaling is still being pursued. While such lithography will eventually lead the way to the theoretical limit for CMOS technology, obstacles such as power and cost are already evident. To see how they arise, it is instructive to return to the scaling of CMOS in theory and in practice to review the primary benefits that have accrued. A discussion of the relation of power and performance to CMOS technology follows.

It is convenient to categorize power into two types—active and passive. This can be accomplished empirically; the power of an integrated circuit (IC), for a fixed operating voltage and temperature, increases linearly with the clock frequency f (the frequency of a master signal with which all operations must be synchronized), driving the IC. Extrapolation of the power vs. frequency response to a frequency of zero (which may be realized in a "sleep" mode) yields a nonzero power, which is referred to as the passive power, $P_{PASSIVE}$. That component of power which is proportional to the frequency is referred to as the active power, P_{ACTIVE}. The active power is due primarily to the charging and discharging of capacitances on the IC, and can be represented by an effective switching capacitance, C_{EFF}, via the well-known relationship

$$P_{ACTIVE} = C_{EFF} V_{DD}^2 f. \quad (2)$$

C_{EFF} does not necessarily represent the actual total capacitance being switched by the chip, since many of the circuits may be switching at some fraction of f (or, for that matter, at some multiple of f). Furthermore, another source of active power, sometimes referred to as "short-circuit," "shoot-through," or "cross-over" power, is also lumped into C_{EFF}. This short-circuit power is due to current which completes a path from the power-supply node to ground directly through a set of n-type and p-type FETs during the short but finite time interval when the gates are close to $V_{DD}/2$, and hence both n- and p-type FETs are in a conducting state. Typically this component represents several percent of the active power.

The passive power can be further refined to two subcategories, one due to circuit design and one due to parasitic leakages that are driven by process technology. Circuit-driven passive power may spring from analog circuits, such as class-A amplifiers, phase-locked loops, and other specialized circuits. These are entirely design-driven and can be managed by suitable design and application architectures. The process-technology passive power consists of the many parasitic currents associated with the device structures, such as junction leakage, gate-

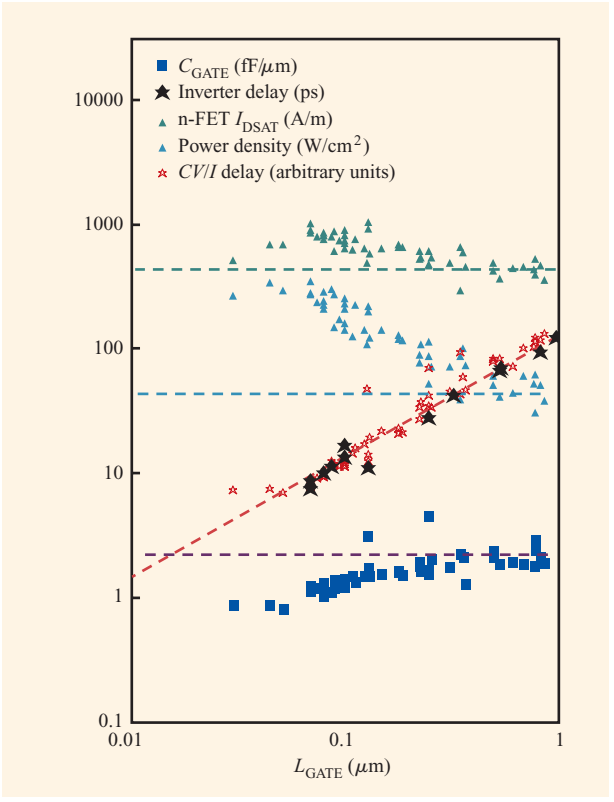

Figure 3

Electrical consequences of industry-trend scaling (points) are contrasted to classic scaling (dashed curves). While the inverter delay scales nearly the same as the classic case (as L_{GATE}), the active-power density does not. Instead, the active-power density increases with decreasing L_{GATE} because of the lag in V_{DD} reduction (see Figure 1), which is only partially mitigated by a reduction in G_{GATE}. I_{DSAT} is the drain current drawn with the gate and drain voltages set at the nominal power-supply voltage, V_{DD}.

induced drain leakage, subthreshold channel currents, gate–insulator tunnel currents, and leakages due to defects. Of these, two are fundamental to the scaling of the technology: the gate–insulator tunnel current and the subthreshold channel current. The gate–insulator tunnel current is due to the quantum-mechanical tunneling of carriers from the gate electrode to the channel (and body) of the FET and has become significant as gate oxide has been thinned to less than 2.8 nm in the 180-nm CMOS generation. Intensive efforts are under way to identify and implement a replacement material for silicon dioxide as the gate dielectric to significantly reduce these tunnel currents. Unfortunately, subthreshold leakage is not susceptible to attack by means of new materials; it remains perhaps the most fundamental challenge facing the VLSI community as scaling proceeds to the 100-nm node and beyond, as explored further below.

The inverter delay, defined as the time required to propagate a transition through a single inverter driving a second, identical inverter, is commonly used as a means of gauging the speed of CMOS transistors (the speed of switching being inversely proportional to the circuit delay). It has been found empirically that a delay, τ, calculated from

$$\tau = C_{GATE}V_{DD}/I_{DSAT} \quad (3)$$

correlates quite well with actual inverter delays. For 100-nm-gate-length n-type MOSFETs, τ typically ranges from 1.5 ps to 3 ps, and about twice as much for p-type, with corresponding inverter delays ranging from 10 ps to 20 ps.

For simplicity, traditional scaling results continue to be used as a framework in which to examine the industry data, even in view of the already noted deviations from such scaling. **Figure 3** illustrates the expected classic scaling consequences, along with data points calculated from the industry scaling trends for I_{DSAT}, C_{GATE}, inverter delay, calculated delay, τ, and switching power density (derived from the product of the power, as described above, and the density). While in "classic scaling" both I_{DSAT} and C_{GATE} remain constant (normalized per MOSFET unit width), the industry-trend data, spanning an L_{GATE} reduction from 1 μm to 100 nm, indicate that I_{DSAT} has nearly doubled. The increase in I_{DSAT} is driven largely by subscaling of V_{DD}; similarly, C_{GATE} has decreased significantly in this period, since L_{GATE} drops more rapidly than T_{OX}, as discussed earlier. As a result, the inverter delay continues to decrease in proportion to (or perhaps slightly faster than) L_{GATE}, as in classic scaling. The switching-power density, $P_{SW} \sim C_{GATE}V_{DD}^2/\tau$, remains constant with classic scaling; hence, the total die switching power shrinks as the decrease in circuit area, $\sim L_{GATE}^2$, thereby allowing more function to be incorporated on a given area of silicon at no increase in switching power. Unfortunately, in contrast to this result, P_{SW}, as calculated from the industry-trend data, has increased by nearly a decade. In this instance, the deviation of the V_{DD} trend from classic scaling has outweighed that of C_{GATE}, to yield this undesirable result. Thus, if die size is kept constant, to add more function with scaling the overall switching power must increase unless some other actions are taken.

Thus, three important benefits arise from classic scaling:

1. Frequency increases $\sim 1/L_{GATE}$; allows faster circuits.
2. Chip area decreases $\sim L_{GATE}^2$; allows reduced-cost circuits.
3. Switching power density $\sim$ constant; allows lower power or more circuits at same power.

As we have already seen, the actual industry data results in modification of only the third benefit; since V_{DD} has not been decreasing as fast as L_{GATE}, the power density has,

in fact, been growing. We will see that this ties strongly into another power-related challenge with scaling, that of passive power.

The Gordian knot of CMOS scaling

A fourth consequence of classic scaling is rather undesirable, but until recently it has not been a particularly negative feature; the standby current density increases exponentially as the length scale is decreased. This follows from the demand that V_T decrease with V_{DD}, together with the observation that $I_{OFF} \sim \exp(-V_T Qe/nkT)$, where Qe is the electronic charge, k is Boltzmann's constant, and T is the absolute temperature. This I_{OFF} dependence is simply a thermodynamic relationship describing the minority-carrier population (the inversion channel) as a function of temperature and energy level in the silicon. While $n \sim 1.4$ for practical designs today, the theoretical lower bound for any FET, even decreasing n to 1, provides only minor reductions to I_{OFF}, given the low values of V_T (~0.2 V) at present. Furthermore, in the most recent generations of CMOS, the rate of tunneling of electrons and holes through gate oxides has increased to a point at which these currents must also be considered. These currents cause an additional power demand in the operation of CMOS which is often referred to as "passive" power, since, unlike switching, or active power, passive power is dissipated by all CMOS circuits all of the time, whether or not they are actively switching.

Figure 4 illustrates the passive-power trend based on subthreshold currents calculated from the industry trends of V_T, all for a junction temperature $T_J = 25°C$. More practical values of T_J only serve to exacerbate this situation, with the off-current of MOSFETs rising nearly two times for each 10°C increase in T_J. For reference, the active-power density shown in Figure 2 is copied onto this scale to illustrate that the subthreshold component of power dissipation is emerging to compete with the long-battled active-power component for even the most power-tolerant, high-speed CMOS applications.

Thus, as the lithography pushes forward, the device designer and the product designer must devise new strategies to cope with the interference of passive power, which pushes for higher V_T (and thus higher V_{DD}) versus active power, which demands lower V_{DD} and thus lower V_T. This results in fragmentation of device design points that address these conflicting needs in the foundry-CMOS business [5, 6], where multiple values of T_{OX}, V_T, L_{GATE}, and V_{DD} are offered within a lithography generation (see **Table 1**). This approach allows the product designer flexibility to choose the best device match for active and passive power vs. performance. Products that are very sensitive to passive power, such as portable and hand-held devices, may sacrifice some performance to enable higher V_T. If these designs require higher performance, they are forced to sacrifice some switching power by use of correspondingly higher V_{DD} as well. Other applications

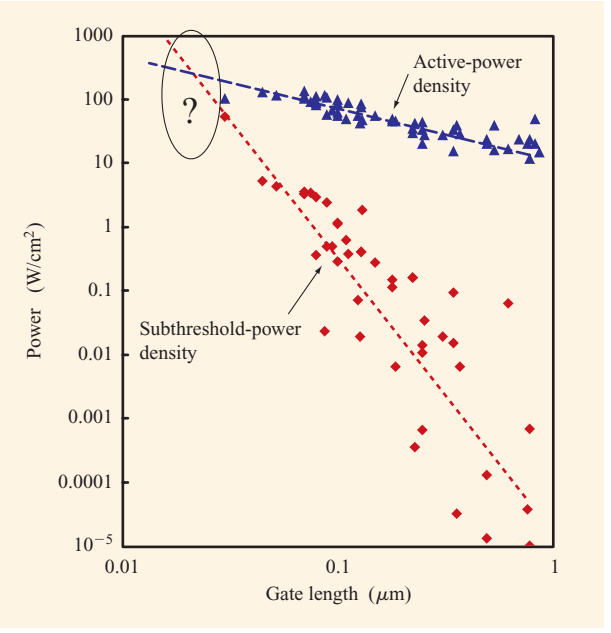

Figure 4

Active-power density and subthreshold-leakage-power density trends calculated from industry trends in Figure 1 are plotted vs. L_{GATE} (points), for a junction temperature of 25°C. Empirical extrapolations (dashed curves) suggest that subthreshold power will equal active power at $L_{GATE} = 20$ nm; this point is encountered closer to $L_{GATE} = 50$ nm when elevated temperatures, typically required of applications, are factored in. This collision, already encountered by applications that are more power-sensitive, will spur further circuit and technology design efforts to manage subthreshold leakages.

Table 1 Foundry CMOS has already been forced to offer a variety of MOSFETs tailored to the demands of individual applications, as illustrated by this variety of devices offered within a 180-nm CMOS technology (after L. K. Han et al. [5]). Where low power, both active and passive, is required, V_{DD} is kept low, T_{OX} high, and V_T high (low $I_{D\text{-}OFF}$). High-performance applications must limit V_{DD} because of active-power density restrictions (cooling), but can afford considerable subthreshold and gate leakage current. Between these cases, one finds general logic with moderate leakage allowances and moderate performance demands.

Application	*High performance*	*1.2-V logic*	*1.5-V logic*	*Low power*	*Interface*
V_{DD} (V)	1.2	1.2	1.5	1.2	2.5
T_{OX} (nm)	1.8	2.2	2.2	2.2	5
$I_{D\text{-}OFF}$ (nA/μm)	10	3	6	0.05	0.01

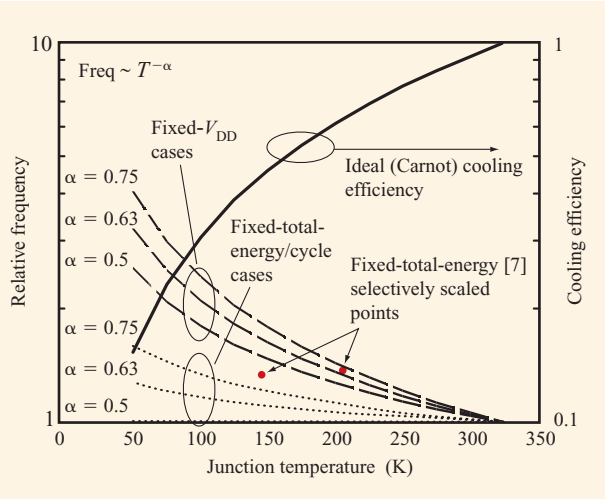

Figure 5

The frequency of a CMOS circuit is (empirically) taken to improve as $T^{-\alpha}$, where T is the absolute temperature and $0.5 < \alpha < 0.75$; a (dashed curves) is dependent on the mechanism-limiting performance (e.g., interconnect vs. transistor delay). As T is decreased (below ambient), power is consumed by the chilling apparatus, ideally described by the Carnot efficiency; here cooling efficiency is defined as the power the chip dissipates divided by the total power required to run and cool the chip (right-hand scale). To maintain fixed total energy per switch (represented by the dotted set of delay curves), V_{DD} is decreased as needed in order to keep the total energy fixed. To avoid this limitation, two "selectively scaled" design points which were demonstrated by Pham [7] enable V_{DD} reduction with less sacrifice in performance by simultaneously decreasing V_T and T_{ox}.

may be challenged to inexpensively conduct heat generated by active power away from the integrated circuits and thus favor lower-V_{DD} devices with low V_T and higher passive power. Thus, the variety of threshold voltages and power-supply voltages offered in 130-nm technology has expanded to address these diverse needs.

Two directions have emerged that offer further specialization in this new era; they are discussed in the following sections.

1. Low temperature for performance-dominated applications

One possible way to avoid the subthreshold-power vs. active-power box may be provided by lower junction temperature. Since I_{OFF} decreases exponentially with $-1/T$, the threshold voltage can be lowered in proportion to T while maintaining constant I_{OFF}, allowing further V_{DD} reduction; temperature cuts the Gordian knot among performance, passive power, and active power. Reduced operating temperature further benefits CMOS performance as a result of increased electron and hole mobilities in MOSFETs, and decreased interconnect resistances. The improvement of performance vs. temperature will depend to some degree on details of the CMOS technology and the product design, since the MOSFET performance can improve as much as T^{-1} to $T^{-0.5}$ depending on process and operating electric field details, while interconnect (resistive) performance may be improved by as much as $T^{-1.5}$. In **Figure 5**, the frequency of the circuit, for a fixed power-supply voltage, will improve as $T^{-\alpha}$, with cases shown for $\alpha = 0.5$, 0.63, and 0.75 to allow for some variability with application. Cooling to 100 K (−172°C) gains two generations of performance (taking $\alpha = 0.63$) and thus looks quite attractive at first glance.

Unfortunately, the process of cooling the circuitry itself requires power, which is proportional to the power dissipated by the circuitry. The Carnot efficiency (energy to run the circuit divided by the sum of the energy to cool the circuit and the energy to run the circuit, with ideal refrigeration) vs. temperature is shown in Figure 5 for the case in which the refrigerator has a heat reservoir at 22°C. This extra power required for cooling must be considered against alternative uses of added power, such as for more parallelism in the circuitry in order to improve computation throughput, or to increase the raw technology speed (e.g., by lower V_T or higher V_{DD}). Then, to make a fair comparison of the benefits of cooling to performance, a second set of frequency vs. temperature loci are shown in Figure 5, where V_{DD} is lowered until the total energy of the circuits plus the refrigerator is equal to the original (room-temperature) value. For this exercise the frequency was taken to be proportional to V_{DD}. The total energy required, then, is taken as the intrinsic CMOS switching energy (fCV_{DD}^2) divided by the Carnot efficiency ($T_{chilled}/T_{ambient}$). The gains in performance obtained for constant voltage with decreased temperature are seriously eroded when constrained by a constant total-power-delay product, with no gain evident for the case $\alpha = 0.5$. Furthermore, the cooling efficiency of real refrigerators is significantly worse than the Carnot efficiency, and it thus becomes readily apparent that cooling is not likely to provide a successful strategy where there is a constraint on total power. It must be remarked, however, that cases exist in which total power is not the relevant limit for the system, and in such cases constant-V_{DD}-constrained performance may be achievable. But, even in these cases, one is frequently bound by other constraints such as the cost or physical volume of the entire computing package, and the benefits must be assessed against alternative strategies such as adding multiple processors or memory, or other features.

Figure 5 shows a novel case of cooling for performance under a total-power constraint, which was experimentally demonstrated by Pham [7] with a PowerPC 603* processor. Based on room-temperature selective scaling

[8], the premise adopted was that for a given manufacturing lithography generation, gains could be made by reducing power within the constraints placed by a fixed lithographic scale. This can be achieved by using refrigeration of CMOS circuits. A 0.5-μm CMOS technology was selectively scaled by reductions of gate dielectric thickness and by reduction of the threshold voltage as a function of temperature. This achieved a rapid reduction of V_{DD} and, therefore, a reduction of the active energy when the temperature was reduced. The gate length was explicitly held fixed at 0.3 μm. As can be seen in Figure 5, a 40% improvement in operating frequency [roughly equivalent to that expected from a generation of CMOS scaling (0.7× delay)] was demonstrated by reducing the temperature by 100°C and scaling V_T and T_{OX} simultaneously. A unique aspect of low-temperature selective scaling is the opportunity for introduction of very-high-k gate dielectric that might not normally find a place in scaling, since the electrically effective value of T_{OX} can be reduced by the use of materials having very high dielectric constants. This is because the problems associated with 2D effects, raised by Frank et al. [9], when using physically thick gate insulators and very-high-k dielectrics, do not arise, since the decrease in effective T_{OX} is not being used to achieve a reduction in L_{GATE}, but rather a reduction in V_{DD}. This opens the possibility for further extensions of this technology direction with very-high-k dielectrics.

Thus, for applications in which frequency is of the utmost importance, and total power and physical volume constraints are relaxed, we see that there is a niche where cooling of CMOS circuits can provide a system performance benefit along the bounds of constant V_{DD}, as shown in Figure 5. However, where total power is constrained, cooling will, at the very least, require process technology changes to realize gains at fixed power.

2. Massive integration with ultralow power

Another direction one could pursue in an attempt to reverse or at least moderate the growing power-density trend shown in Figure 4 is to minimize the energy spent per operation by the use of very low V_{DD}. When V_{DD} is lowered much more rapidly than the extrapolated trend, large circuit counts become attainable at reasonable power budgets, and (presumably) one can then achieve system performance through massive parallelism. Pushing this idea to the extreme, the lowest operating voltages are achieved when MOSFETs are operated entirely in the subthreshold regime: $V_{DD} < V_T$. Subthreshold-operated inverters have been experimentally demonstrated to operate on V_{DD} as low as 70 mV at room temperature [10], compared to the theoretical minimum for V_{DD} with bistable logic states, which been shown to range from 36 mV to ~80 mV, depending on circuit details and

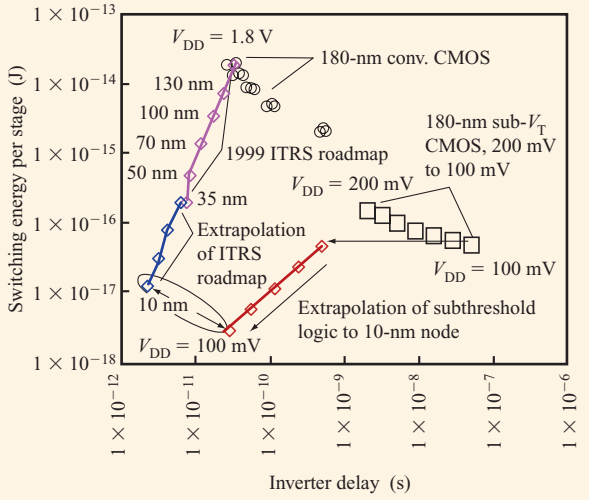

Figure 6

Energy–delay products for standard CMOS logic and for subthreshold logic are examined; the 1999 ITRS roadmap is extended from the measured 180-nm conventional CMOS data point. Extrapolations are made from the ITRS roadmap (ending at 35 nm) down to a hypothetical 10-nm node as well for comparison to subthreshold CMOS. The 180-nm subthreshold operating point at V_{DD} = 100 mV is also extrapolated, first to a design with MOSFETs at V_T = 0.1 V, and then to a 10-nm subthreshold-logic node. A "natural" convergence is seen, as suggested earlier in Figure 4.

MOSFET characteristics [11, 12]. **Figure 6** shows a comparison of energy–delay product for conventional 180-nm-generation CMOS logic and for an experimental 180-nm-generation subthreshold logic [13], with V_{DD} varied between 100 mV and 200 mV. Stage delay is used as a dependent variable to illustrate the tradeoff available between speed and energy per logic operation. While the lowest voltage at which these inverters can remain operable is 36 mV at room temperature (given an ideal subthreshold swing of 60 mV/decade at 25°C), in practice one must require operation of at least two input NANDs or two input NORs in order to accomplish useful computations. Also, to allow for some tolerance to process and design margins, operation at V_{DD} = 100 mV may prove a practical lower bound.

Of course, the very nature of subthreshold CMOS requires very good matching between FET threshold voltages (more precisely, matching of the off-currents, since this is what limits bistability of subthreshold CMOS circuits). In particular, n-FETs and p-FETs must be very well matched to one another, and this requirement is most demanding in that many parts of a CMOS process may introduce significant independence between the n-FET

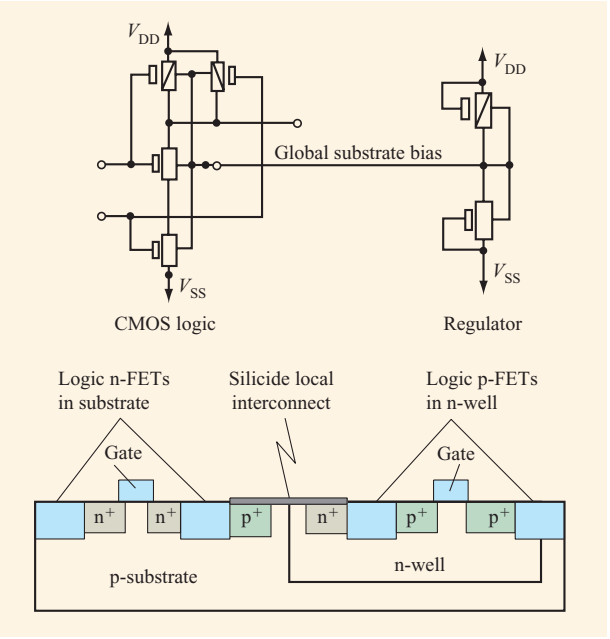

Figure 7

Subthreshold logic requires precise matching of n-FET and p-FET subthreshold currents, despite the fact that the processes which determine these currents have considerable independence. This problem is avoided by the simple regulator circuit shown, which connects the n-well to the substrate and biases the combination to a bias suitable for subthreshold operation. Reprinted with permission from [13]; © 2001 IEEE.

and p-FET threshold voltages. A solution to this problem has recently been provided [13] by locally connecting n-wells of p-type FETs to the global substrate in a p-substrate CMOS technology, and then biasing the substrate to a voltage which matches the n-FETs to the p-FETs. A simple body-driven operational amplifier, shown in **Figure 7**, will arrive at a body bias at which the off-currents of n-FETs and p-FETs are matched, ensuring functional operation of the subthreshold logic on the die. This reduces the matching requirements to intra-die matching of like FETs. Body doping fluctuations, which give rise to random threshold variations, could prohibit the scaling of this scheme to small dimensions for ordinary bulk-controlled MOSFETs; however, back-gated MOSFETs with bodies of nearly intrinsic silicon, and V_T set by back-gate bias, avoid this limitation and could enable scaling of subthreshold logic to scales approaching the atomic level.

An interesting exercise is also illustrated in Figure 6. The "on" current density in [13] was set at ~20 nA/μm, but by choice of lower V_T, the "on" current could be increased to 2000 nA/μm while remaining in subthreshold conduction. This would decrease the delay for the V_{DD} = 100-mV case by two decades without substantially changing the energy per cycle, as indicated by the 180-nm extrapolation point in Figure 6. Scaling of this design point results in performance increasing as $1/L_{GATE}$ (driven by subthreshold current increasing as $1/L_{GATE}$), while the energy decreases as L_{GATE}, as indicated in Figure 6 by the subthreshold scaling extrapolation curve, down to a 10-nm node. For comparison, the ITRS projections were extrapolated to a 10-nm node. The energy–delay curves for standard CMOS and subthreshold CMOS converge as the 10-nm node approaches, and the 10× difference in energy between these two cases is simply driven by a projected 300-mV supply voltage for the ITRS case vs. 100 mV for the subthreshold case (recall that energy is quadratic in V_{DD}). Thus, one sees that seemingly highly disparate device design strategies converge to very similar points as we travel down the last leg of the scaling path.

In summary, the convergence of subthreshold power with active power for conventional high-speed CMOS (Figure 4) and the convergence of a 10-nm-extrapolated ITRS roadmap with a 10-nm-extrapolated subthreshold CMOS suggest that a low-power path, possibly invoking subthreshold techniques for power reduction, is very likely to play a role on the scaling path to the 10-nm node, as active-power constraints drive further innovation. Continued advances in chip and software architecture may well harness massive parallelism, and employment of CMOS directions similar to the subthreshold approach could prove to be fruitful in navigating the power flood presented when scaling below the 50-nm node.

Scaling toward the atomic limit

The present microelectronic technology progresses by advances in lithographic capabilities. Up to this point, revolutionary departures from planar CMOS could not compete with rote scaling: The benefits from scaling have always provided very rigorous standards against which any new structures, requiring extraordinary development and exploratory research, simply could not compete. But in an era in which rote scaling has been halted, or at least radically impeded, these alternative approaches may provide the most effective path to achieving improved power, performance, and density.

The first improvement addressed is one of device architecture. IBM has already pushed forward from conventional bulk MOSFETs to SOI MOSFETs [14], in recognition of the impending need for greater extendibility in the scaling of transistors. Others [15–17] have also proceeded toward SOI, although pursuit of scaling in conventional bulk CMOS [18] persists. Purely from the point of view of device architecture, there is a widely held opinion that double-gate MOSFETs (or double-gate CMOS: DGCMOS) provides the ideal structure for scalability [19]; what remains hotly debated is whether

a manufacturable realization of this architecture can be developed. While the principles of DGCMOS are covered in greater depth elsewhere in this issue, three observations merit explicit mention here:

1. All MOSFETs control short-channel effects (SCE) by maximizing gate and minimizing drain coupling to the channel; the latter is accomplished, in bulk and in partially depleted SOI devices, by shielding the channel from the drain by minimizing the distance between the neutral (electrically conductive) region of the body (e.g., the body is heavily doped) and the (surface) channel. Thus, higher body coupling to the channel results in better control of short-channel effects. This directs development toward small body depletion depth.
2. Drive current is determined by the product of the charge-carrier density of the channel and the velocity of those carriers. Increased body coupling to the channel intrinsically reduces the channel charge density for a given gate coupling, since the gate coupling must compete directly with this body coupling. Thus, higher body coupling to the channel results in lower drive current. This directs development toward large body depletion depth.
3. Once L_{GATE} and T_{OX} are fixed (usually by manufacturing constraints and power-supply voltage), the device designer is left with the tradeoff of increasing body coupling to the channel (reducing body depletion depth) until adequate SCE is achieved, while avoiding excessively large body coupling, which would result in reduced drive current. Drive current must be compromised in order to achieve good V_T control; as V_{DD} is reduced to further enable CMOS scaling, V_T control must improve, and this compromise becomes more acute.

The DG MOSFET resolves the conflict between controlling short-channel effects and maximizing drive current, since control of body coupling to the drain is shifted from the (charge-neutral) body to a back gate, which is also driven by the input signal. Hence, one wins by increasing the coupling of the back gate, both in short-channel control and in drive.

Naturally the question arises as to why DGCMOS is not in manufacturing today. The answer becomes clear on inspection of the required structure, illustrated schematically in **Figure 8**. A back gate must be self-aligned with the source and drain junctions, as well as the front gate, in order to avoid highly penalizing parasitic capacitances. Furthermore, the back gate must be connected via a low-resistance path to the front gate, and it must have low parasitic capacitances to other technology elements present, such as the wafer substrate, source, and drain, in order to avoid substantial performance (and

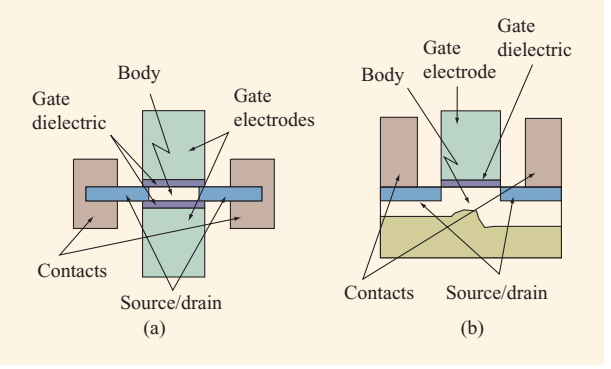

Figure 8

(a) Double-gate MOSFET compared to (b) conventional single-gate MOSFET. The single-gate device has source, drain, and gate completely accessible from the surface of the silicon wafer, facilitating manufacturable, self-aligned processes. The double-gate MOSFET poses the challenge of adding the second (bottom) gate in a manner that provides self-alignment to both the top gate and the source and drain regions.

power) degradation. Many schemes to achieve efficient DG MOSFETs have been proposed, but recently the FinFET [20–22], an improved version of the delta device [23], has begun to show great promise in enabling entry of DGCMOS to CMOS manufacturing. This device structure provides for DGCMOS devices constructed with conventional planar manufacturing processes, while satisfying the requirements of multiple self-alignments and low gate resistance, by literally turning the silicon channel on its side, yielding access to both the "front" and "back" gates from the top of the wafer during processing. This makes self-alignment of the gates with one another, and alignment of the source and drain regions with both gates, relatively straightforward, and it is also compatible with access to both gates through relatively low-resistance paths. **Figure 9** is a schematic illustration of a FinFET with an inset SEM micrograph of a prototype n-MOS FinFET fabricated at IBM. Well-behaved series resistance and gate-to-drain capacitances, as well as CMOS FinFETs with V_Ts compatible with sub-one-volt CMOS logic, have recently been demonstrated at IBM [24], and this structure is likely to challenge the purely planar MOSFET structures for dominance in the ULSI technology in the not-too-distant future.

The FinFET presents some unfamiliar physical characteristics which are worth discussion. The MOSFET width in conventional devices is associated with the drive of the device and is varied by making the planar silicon island wider. In a FinFET the effective width is determined by (up to twice) the height of the fin (see Figure 8). Larger effective widths are achieved by adding

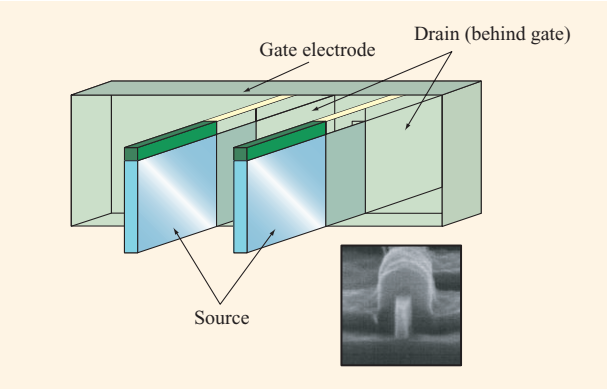

Figure 9

Schematic illustration of a FinFET; inset is a SEM micrograph of a prototype n-MOS FinFET. The FinFET is an attractive candidate for the realization of DGCMOS, since both gates are immediately accessible from the top of the wafer during processing, as are both sides of the source/drain electrodes. Conventional design flexibility is preserved, with the number of fins increasing in steps set by the design width of the transistor. It is easily seen that the net silicon channel surface available to the circuit designer is greater than that provided by planar transistors, provided that the fin height exceeds one-half the fin pitch (it is presumed that both gates have the same work function). Further gains in current density are, in principle, available through increasing the "height" of the silicon fins, adding a new dimension to device scaling.

many "fins" in parallel to provide larger drive current when called for. Channel length is determined, as in conventional planar CMOS devices, by the length of the gate electrode, which is horizontally defined by lithographic means. The body (fin) thickness must be approximately one-fourth of the length of the gate (or less) to ensure suppression of deleterious short-channel effects, such as variability in threshold voltage and excessive drain leakage currents. It is essential that the gate length itself be designed at the minimum physically achievable size, given the current lithographic capability, in order to minimize gate capacitance and maximize device drive; hence, it follows that the thickness of the fin must be defined by some means beyond that available from conventional lithography. Sidewall image transfer (SIT) provides one means of achieving sublithographic fins with the required dimensions and control. In SIT [25], normal lithographic techniques are used to pattern a mandrel material, which is then used to form a spacer on the edge of the mandrel. After removal of the mandrel, the remaining spacer acts as a mask whose size and tolerance are determined by deposition and etch tolerances. Tolerances between 5% and 10% can readily be engineered for such processes at sizes well below those accessible to lithography. For example, sidewall spacers have been manufactured at 25% of minimum image size with excellent control for many years in conventional CMOS technologies.

What can be expected to follow when silicon technology approaches the atomic limit, and structures on silicon have dimensions of the order of several nanometers? Perhaps radical departure from CMOS, such as nanotube or molecular switches, will provide new directions, or it is entirely possible that no successor to CMOS will appear. Should the latter prove to be the case, an intriguing possibility is presented by the observation that the device widths in the FinFET architecture can be increased at a fixed lithographic scale by increasing the height of the silicon fins, thus providing more device area in a physical area than is possible with planar devices. While the MOSFET performance as measured by CV/I delay is not improved, since both C_{GATE} and I_{DSAT} increase in direct proportion to the fin height, interconnect contributions to delay may be decreased by allowing for closer placement of MOSFETs of the same drive capability and hence lower interconnect capacitance and resistance. And such interconnect delays already present major obstacles to scaling CMOS designs. Thus, one new direction (literally) for device scaling could become the vertical direction with respect to the wafer plane. Carrying this idea even further, we can envision even more vertical integration and perhaps can expect that this may require operation in some low-power-enabled mode such as multiple-threshold CMOS [26] or subthreshold CMOS. In any case, the economic incentives for improved density, lower power, and higher performance will find a new era of technology innovation. A most difficult question to answer will then be this: "When will one's daily encounter with silicon-based devices be as specialized as an encounter with vacuum electronics is today?"

Summary and conclusion

It is argued that after more than two decades of CMOS scaling, we are now entering the first of two significant transitions that will occur in the CMOS ULSI arena, namely the era of increased device specialization by application. Further expansion in the specialization of device structures and design points will proceed over the next few generations of CMOS, with increased emphasis on new materials and structures; this will maintain the momentum toward power, performance, and cost benefits that, until recently, had been simply benefits of scaling. Beyond this first transition, a point will be reached at which further gains from scaling of traditional planar CMOS devices will be very difficult, limited by leakage and switching power considerations. At this point, the planar MOSFET will be challenged by 3D or nearly 3D structures that are amenable to planar fabrication techniques without disruption of CMOS fabrication

facilities; the FinFET currently provides the most likely candidate for succession, enabling continued growth in density and reduction of cost for ULSI circuits, even as the industry approaches the second transition nearing the atomic limit.

Most excitingly, we are approaching the end of an era of scaling gains by rote shrinkage of device dimensions, and entering a post-scaling era, a new phase of CMOS evolution in which innovation is demanded simply to compete. The trends in benefits to density, performance, and power will be continued through such innovations. Rather than coming to a close, a new era of CMOS technology is just beginning.

Acknowledgments

I am most indebted and grateful to many colleagues and coworkers for stimulating discussions on the future of CMOS logic. I particularly thank Peter Cottrell for many in-depth discussions of power tradeoffs and design points, and for distilling many thoughts that entered into this work. I also am grateful to Paul Solomon for suggesting some of the topics addressed here and for his insights and helpful thoughts.

References

1. *International Technology Roadmap for Semiconductors*, 1999 Edition, Semiconductor Industry Association, 4300 Stevens Suite Boulevard, Suite 271, San Jose, CA 95129.
2. H. Ooka, S. Murakami, M. Murayama, K. Yoshida, S. Takao, and O. Kudoh, "High Speed CMOS Technology for ASIC Application," *IEDM Tech. Digest*, pp. 240–243 (1986).
3. S. Song, J. H. Yi, W. S. Kim, J. S. Lee, K. Fujihara, H. K. Kang, J. T. Moon, and M. Y. Lee, "CMOS Device Scaling Beyond 100 nm," *IEDM Tech. Digest*, pp. 235–238 (2000).
4. R. Dennard, F. Gaensslen, H. Yu, V. Rideout, E. Bassous, and A. LeBlanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits* **SC-9,** 256–268 (1974).
5. L. K. Han, S. Biesemans, J. Heidenreich, K. Houlihan, C. Lin, V. McGahay, T. Schimi, A. Schmidt, U. P. Schroeder, M. Stetter, C. Wann, D. Warner, R. Mankopf, and B. Chen, "A Modular 0.13μm Bulk CMOS Technology for High Performance and Low Power Applications," *IEEE Symposium on VLSI Technology*, *Digest of Technical Papers*, 2000, pp. 12–13.
6. C. H. Diaz, M. Chang, W. Chen, M. Chiang, H. Su, S. Chang, P. Lu, K. Pan, C. Yang, L. Chen, C. Su, C. Wu, C. H. Wang, C. C. Wang, J. Shih, H. Hsieh, H. Tao, S. Jang, M. Yu, S. Shue, B. Chen, T. Chang, C. Hou, B. K. Liew, K. H. Lee, and Y. C. Sun, "A 0.15 μm CMOS Foundry Technology with 0.1 μm Devices for High Performance Applications," *IEEE Symposium on VLSI Technology*, *Digest of Technical Papers*, 2000, pp. 146–147.
7. D. Pham, "Selective Device-Temperature Scaling for Optimum Performance–Power in MOSFET Circuit Design," Ph.D. thesis, University of Vermont, Burlington, May 1998.
8. K. Bernstein, J. Bertsch, L. Heller, E. Nowak, and F. White, "Experimental 2.0V Power/Performance Optimization of a 3.6V-Design CMOS Microprocessor-PowerPC 602," *IEEE Symposium on VLSI Technology*, *Digest of Technical Papers*, 1994, pp. 83–84.
9. D. J. Frank, Y. Taur, and H.-S. P. Wong, "Generalized Scale Length for Two-Dimensional Effects in MOSFETs," *IEEE Electron Device Lett.* **19,** 385 (1998).
10. J. B. Burr, "Cryogenic Ultra Low Power CMOS," *Proceedings of the IEEE Symposium on Low Power Electronics*, 1995, pp. 82–83.
11. G. Schrom, C. Pichler, T. Simlinger, and S. Selberherr, "On the Lower Bounds of CMOS Supply Voltage," *Solid State Electron.* **39,** 225–430 (1996).
12. D. J. Frank, R. Dennard, E. J. Nowak, P. Solomon, Y. Taur, and H.-S. Wong, "Device Scaling Limits of Si MOSFETs and Their Application Dependencies," *Proc. IEEE*, pp. 259–288 (March 2001).
13. A. Bryant, J. Brown, P. Cottrell, J. Ellis-Monaghan, M. Ketchen, and E. J. Nowak, "Low Power CMOS at $V_{dd} = 4kT/q$," *Conference Digest, 59th Device Research Conference*, June 2001, pp. 22–23.
14. A. Ajmera, J. W. Sleight, F. Assaderaghi, R. Bolam, A. Bryant, M. Coffey, H. Hovel, J. Lasky, E. Leobandung, W. Rausch, D. Sadana, D. Schepis, L. F. Wagner, K. Wu, and B. Davari, "0.22 μm CMOS–SOI technology with a Cu BEOL," *IEEE Symposium on VLSI Technology*, *Digest of Technical Papers*, 1999, pp. 15–16.
15. K. Cox, J. Scott, S. Bishop, M. Bhat, B. Mettleton, D. Pan, M. Hamilton, D. Chang, L. Day, and P. Schani, "A Partially Depleted 1.8V SOI SRAM Technology Featuring a 3.77μm^2 Cell," *IEEE Symposium on VLSI Technology*, *Digest of Technical Papers*, 2000, pp. 13–15.
16. K. Sukegawa, M. Yamaji, K. Yoshie, K. Furomochi, T. Maruyama, H. Morioka, N. Naori, T. Kubo, H. Kanata, M. Kai, S. Satoh, T. Izawa, and K. Kubota, "High-Performance 80-nm Gate Length SOI–CMOS Technology with Copper and Very-Low-*k* Interconnects," *IEEE Symposium on VLSI Technology*, *Digest of Technical Papers*, 2000, pp. 186–187.
17. M. Ino, H. Sawada, K. Nishimura, M. Urano, H. Sato, S. Date, T. Ishihara, T. Takeda, Y. Kado, H. Inokawa, T. Tsuchiya, Y. Sakakibara, Y. Arita, K. Izumi, K. Takeya, and T. Sakai, ".25μm CMOS/SIMOX Gate Array LSI," *IEEE International Solid-State Circuits Conference*, *Digest of Technical Papers*, 1996, p. 86.
18. R. Chau, J. Kavalieros, B. Roberds, R. Schenker, D. Lionberger, D. Barlage, B. Doyle, R. Arghavani, A. Murthy, and G. Dewey, "30nm Physical Gate Length CMOS Transistors with 1.0 ps n-MOS and 1.7 ps p-MOS Gate Delays," *IEDM Tech. Digest*, pp. 45–48 (2000).
19. S. Christoloveanu, T. Ernst, D. Munteanu, and T. Ouisse, "Ultimate MOSFETs on SOI: Ultra Thin, Single Gate, Double Gate, or Ground Plane," *Int. J. High Speed Electron. & Syst.* **10,** 217–230 (2000).
20. X. Huang, W.-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y.-K. Choi, K. Asano, V. Subramanian, T.-J. King, J. Bokor, and C. Hu, "Sub 50-nm FinFET," *IEDM Tech. Digest*, pp. 67–70 (1999).
21. D. M. Fried, A. P. Johnson, E. J. Nowak, J. Rankin, and C. R. Willets, "A sub-40 nm Body-Thickness N-type FinFET," *Conference Digest, 59th Device Research Conference*, June 2001, pp. 24–25.
22. D. Hisamoto, W.-C. Lee, J. Kedzierski, A. Anderson, H. Takeuchi, K. Asano, T.-J. King, J. Bokor, and C. Hu, "A Folded-Channel MOSFET for Deep-Sub-Tenth Micron Era," *IEDM Tech. Digest*, pp. 1032–1034 (1998).
23. D. Hisamoto, T. Kaga, Y. Kawamoto, and E. Takeda, "A Fully Depleted Lean-Channel Transistor (DELTA)," *IEDM Tech. Digest*, pp. 833–836 (1989).

24. J. Kedzierski, D. M. Fried, E. J. Nowak, T. Kanarsky, J. Rankin, H. Hanafi, W. Natzle, D. Boyd, Y. Zhang, R. Roy, J. Newbury, C. Yu, Q. Yang, P. Saunders, C. Willets, A. Johnson, S. Cole, H. Young, N. Carpenter, D. Rakowski, B. A. Rainey, P. Cottrell, M. Ieong, and P. Wong, "High-Performance Symmetric Gate and CMOS-Compatible Vt Asymmetric Gate FinFET Devices," *IEDM Tech. Digest*, pp. 437–440 (2001).
25. "Method for Making Submicron Dimensions in Structures Using Sidewall Image Transfer Techniques," C. Johnson, S. Ogura, J. Riseman, N. Rovedo, and J. J. Shepard, *IBM Tech. Disclosure Bull.* **02-84,** 4587–4589 (1984).
26. "1-V Power Supply High-Speed Digital Circuit Technology with Multi-Threshold Voltage CMOS," M. Mutoh, T. Douskei, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, *IEEE J. Solid-State Circuits* **30,** 847–854 (1995).

Received June 21, 2001; accepted for publication January 3, 2002

Edward J. Nowak *IBM Microelectronics Division, 1000 River Road, Essex Junction, Vermont 05452 (ejnowak@us.ibm.com).*
Dr. Nowak is a Senior Technical Staff Member in the CMOS Process Development Group. He received a B.S. degree in physics from M.I.T. in 1973, and M.S. and Ph.D. degrees in physics from the University of Maryland in 1974 and 1978, respectively. Following postdoctoral research at New York University, he joined IBM in Essex Junction in 1981 to work on memory device and process development. In 1985 he turned to high-speed CMOS logic devices, spanning 1-μm to 50-nm device designs. Dr. Nowak's current research interests are in double-gate CMOS and low-power CMOS. He is an author or coauthor of more than 50 U.S. patents and more than 40 technical papers and a book in the areas of VLSI technology and design. Dr. Nowak is a member of the Institute of Electrical and Electronics Engineers and the American Physical Society.