CMOS design near the limit of scaling

by Y. Taur

Beginning with a brief review of CMOS scaling trends from 1 μm to 0.1 μm, this paper examines the fundamental factors that will ultimately limit CMOS scaling and considers the design issues near the limit of scaling. The fundamental limiting factors are electron thermal energy, tunneling leakage through gate oxide, and 2D electrostatic scale length. Both the standby power and the active power of a processor chip will increase precipitously below the 0.1-μm or 100-nm technology generation. To extend CMOS scaling to the shortest channel length possible while still gaining significant performance benefit, an optimized, vertically and laterally nonuniform doping design (*superhalo*) is presented. It is projected that room-temperature CMOS will be scaled to 20-nm channel length with the superhalo profile. Low-temperature CMOS allows additional design space to further extend CMOS scaling to near 10 nm.

1. Introduction

The steady downscaling of transistor dimensions over the past two decades has been the main stimulus to the growth of silicon integrated circuits (ICs) and the information industry. The more an IC is scaled, the higher becomes its packing density, the higher its circuit speed, and the lower its power dissipation [1]. These have been key in the evolutionary progress leading to today's computers and communication systems that offer superior performance, dramatically reduced cost per function, and much-reduced physical size compared to their predecessors.

The prevailing VLSI technology today comprises CMOS devices because of their unique characteristic of negligible standby power, which allows the integration of tens of millions of transistors on a processor chip with only a very small fraction ($<1\%$) of them switching at any given instant. As the CMOS dimension, in particular the channel length, is scaled to the nanometer regime (<100 nm), however, the electrical barriers in the device begin to lose their insulating properties because of thermal injection and quantum-mechanical tunneling [2]. This results in a rapid rise of the standby power of the chip, placing a limit on the integration level as well as on the switching speed.

This paper considers bulk CMOS designs that extend the limit of scaling to the shortest channel length possible. Section 2 is an overview of the perspectives of CMOS scaling which examines the fundamental factors that will ultimately limit scaling: electron thermal voltage and oxide tunneling. Section 3 discusses a 2D MOSFET scale length model and presents a feasible design for 25-nm CMOS, likely to be near the limit of CMOS scaling. Section 4 explores the possibility and design issues of extending CMOS scaling to 10-nm channel length through low-temperature operation. Section 5 concludes the paper.

2. CMOS scaling trends and limiting factors

When the dimensions of a MOSFET are scaled down, both the voltage level and the gate-oxide thickness must also be reduced [1]. Since the electron thermal voltage,

0018-8646/02/$5.00 © 2002 IBM

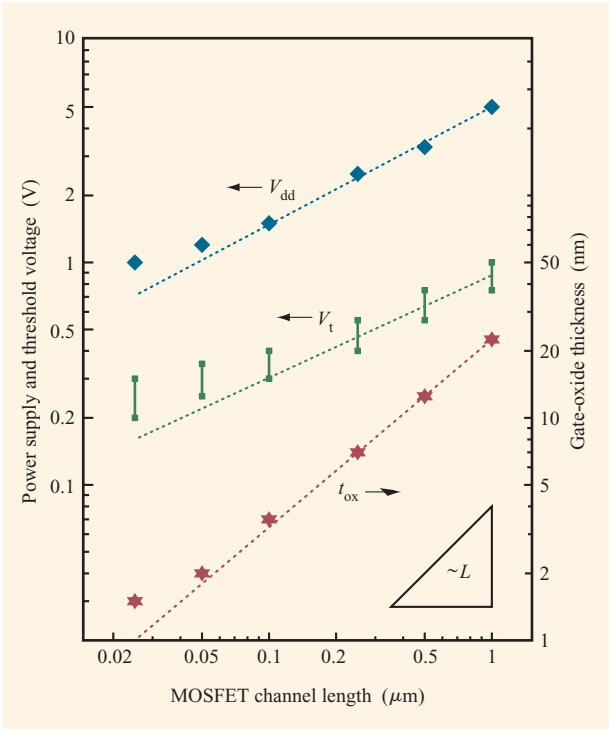

Figure 1

History and trends of power-supply voltage (V_{dd}), threshold voltage (V_t), and gate-oxide thickness (t_{ox}) vs. channel length for CMOS logic technologies.

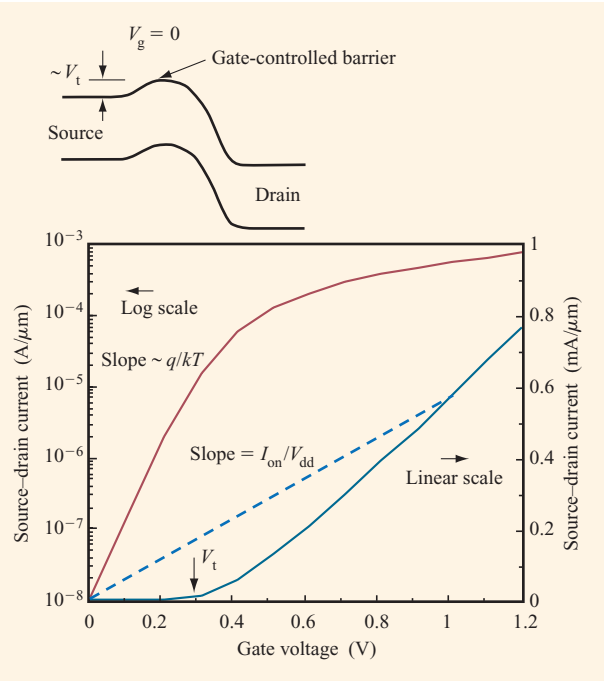

Figure 2

MOSFET current in both logarithmic (left) and linear (right) scales vs. gate voltage. The slope of the dotted line represents the large-signal transconductance for a digital circuit. Inset shows the band diagram of an n-MOSFET. The barrier height at $V_g = 0$ is proportional to V_t.

kT/q, is a constant for room-temperature electronics, the ratio between the operating voltage and the thermal voltage inevitably shrinks. This leads to higher source-to-drain leakage currents stemming from the thermal diffusion of electrons. At the same time, the gate oxide has been scaled to a thickness of only a few atomic layers, where quantum-mechanical tunneling gives rise to a sharp increase in gate leakage currents [3]. The effects of these fundamental factors on CMOS scaling are quantified below.

Power supply and threshold voltage

Figure 1 shows the scaling trend of power-supply voltage (V_{dd}), threshold voltage (V_t), and gate-oxide thickness (t_{ox}) as a function of CMOS channel length [2]. It is seen that the power-supply voltage has not been decreasing at a rate proportional to the channel length. This means that the field has been gradually rising over the generations between 1-μm and 0.1-μm channel lengths. Fortunately, thinner oxides are more reliable at high fields, thus allowing operation at the reduced but nonscaled supply voltages.

Below 0.1 μm, threshold voltage deviates even further from the past scaling behavior, as depicted in Figure 1.

MOSFET threshold voltage is defined as the gate voltage at which significant current begins to flow from the source to the drain (**Figure 2**). Below the threshold voltage, the current does not drop immediately to zero. Rather, it decreases exponentially, with a slope on the logarithmic scale inversely proportional to the thermal energy kT. This is because some of the thermally distributed electrons at the source of the transistor have high enough energy to overcome the potential barrier controlled by the gate voltage and flow to the drain (Figure 2 inset). Such a subthreshold behavior follows directly from fundamental thermodynamics and is independent of power-supply voltage and channel length.

The standby power of a CMOS chip due to source-to-drain subthreshold leakage is given [4] by

$$P_{off} = W_{tot} V_{dd} I_{off} = W_{tot} V_{dd} I_0 \exp\left(-\frac{qV_t}{mkT}\right), \quad (1)$$

where W_{tot} is the total turned-off device width with V_{dd} across the source and drain, I_{off} is the average off-current per device width at 100°C (worst-case temperature), I_0 is the extrapolated current per width at threshold voltage (of the order of 1–10 μA/μm for 0.1-μm devices),

m is a dimensionless ideality factor typically ≈1.2 (to be discussed later), and V_t is the threshold voltage at 100°C. Even if V_t is kept constant, the leakage current of turned-off devices will increase in proportion to $1/t_{ox}$ and (W_{tot}/L) because the current at threshold condition I_0 is proportional to the inversion charge density at threshold, $Q_i \sim (1–2)(kT/q)C_{ox}$ [4], where $C_{ox} = \varepsilon_{ox}/t_{ox}$ is the gate-oxide capacitance per unit area. The off-state leakage current would further increase by about ten times for every 0.1-V reduction of V_t. For a chip with an integration level of 100 million transistors, the average leakage current of turned-off devices should not exceed a few times 10^{-8} A. This constraint holds the threshold voltage to a minimum of about ~0.2 V at the operating temperature (100°C worst case).

The saturation of threshold voltage leads to a saturation of the power-supply voltage as well, at a minimum value of about ~1.0 V, as shown in Figure 1. This is because CMOS performance, measured by the large-signal transconductance I_{on}/V_{dd} (slope of the dotted lines in Figure 2), degrades rapidly with an increasing ratio of V_t/V_{dd}. For high-performance CMOS, a V_t/V_{dd} ratio of <0.3 is desired [4]. A nonscaled V_{dd} means more aggressive device designs at higher electric fields. More significantly, it drives up the active power of a CMOS chip (crossover currents are usually negligible), given by

$$P_{ac} = C_{sw}V_{dd}^2 f, \quad (2)$$

where C_{sw} is the total node capacitance being charged and discharged in a clock cycle, and f is the clock frequency. As CMOS technology advances, clock frequency goes up. The total switching capacitance is likely to increase as well, as one tries to integrate more circuits into the same or an even larger chip area. The active power of today's high-end microprocessors is already in the 50–100-W range. Barring a major breakthrough in power-management systems via architectural innovation, expensive packaging solutions will very soon be required in order to dissipate the heat generated by the chip.

There are other schemes for meeting leakage power requirements. For example, one can fabricate multiple-threshold-voltage devices on a chip [5]. Low-threshold devices could be used in critical logic paths for speed, while high-threshold devices would be used everywhere else, including memory arrays, for low standby power. One can also sense the circuit activity and cut off the power supply to logic blocks that are not switching—an approach known as sleep mode. Other possibilities include dynamic-threshold devices, for which the threshold voltage is controlled by a back-gate bias voltage in either bulk or silicon-on-insulator device structures. Yet another option is low-temperature CMOS. Low-temperature operation not only steepens the subthreshold slope and improves

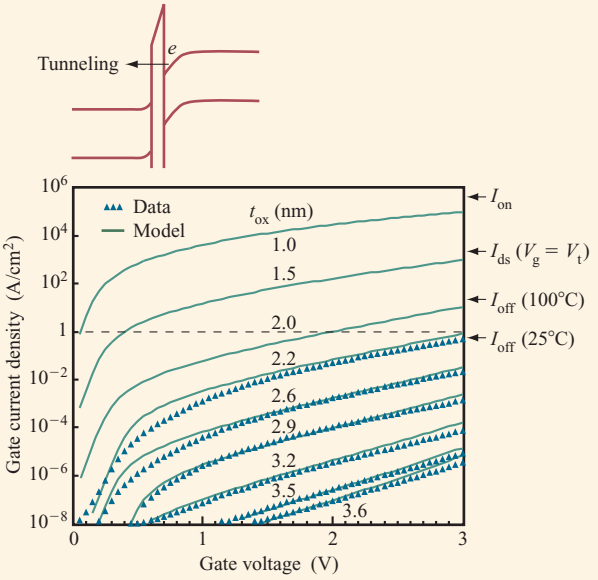

Figure 3

Measured and calculated oxide tunneling currents vs. gate voltage for different oxide thicknesses. Labels on the right, from the bottom up, mark the order of magnitude of off-currents at room and worst-case temperatures, source-to-drain current at $V_g = V_t$ ($V_{ds} = V_{dd}$), and on-current at $V_g = V_{ds} = V_{dd}$. The inset shows the band diagram for tunneling in a turned-on n-MOSFET.

mobility (Section 4), but also reduces wire resistance. However, all of these solutions generally carry a cost in density and complexity.

Gate-oxide tunneling

To keep adverse 2D electrostatic effects on threshold voltage (i.e., short-channel effects) under control (discussed in more detail in Section 3), gate-oxide thickness is reduced nearly in proportion to channel length, as shown in Figure 1. This is necessary in order for the gate to retain more control over the channel than the drain. For CMOS devices with channel lengths of 100 nm or less, an oxide thickness of <3 nm is needed. This thickness comprises only a few layers of atoms and is approaching fundamental limits. While it is amazing that SiO_2 can carry us this far without being limited by extrinsic factors such as defect density, surface roughness, or large-scale thickness and uniformity control, oxide films this thin are subject to quantum-mechanical tunneling, giving rise to a gate leakage current that increases exponentially as the oxide thickness is scaled down. Tunneling currents for oxide thicknesses ranging from 3.6 to 1.0 nm are plotted versus gate voltage in **Figure 3** [3]. In the direct-tunneling regime, the current is rather

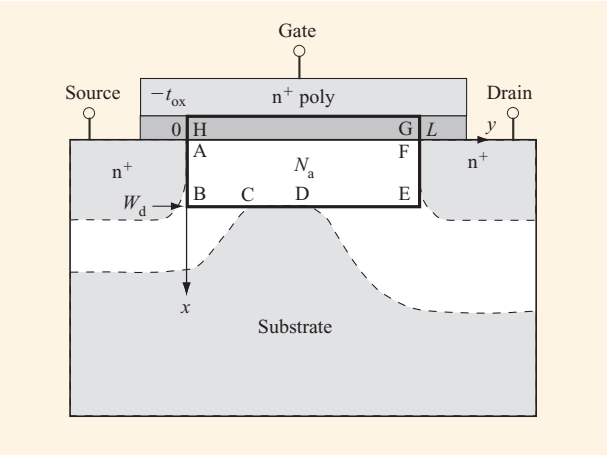

Figure 4

Simplified geometry for analyzing 2D effects in a short-channel MOSFET. The white area in silicon represents the depletion regions where mobile carriers are swept away by the built-in as well as the applied fields of the gate and between the substrate and the source and drain. The solution to the electrostatic potential is reduced to that of a 2D boundary-value problem for the rectangle bounded by heavy dark lines.

insensitive to the applied voltage or field across the oxide, so reduced-voltage operation will not buy much relief. Although the gate leakage current may be at a level that is negligible compared with the on-state current of a device, it will first have an effect on the chip standby power. Note that the leakage power will be dominated by turned-on n-MOSFETs, in which electrons tunnel from the silicon inversion layer to the positively biased gate (Figure 3 inset). Edge tunneling in the gate-to-drain overlap region of turned-off devices should not be a fundamental issue, since one can always build up the corner oxide thickness by additional oxidation of polysilicon after gate patterning. p-MOSFETs have a much lower leakage than n-MOSFETs because there are very few electrons in the p^+ polysilicon ("poly") gate available for tunneling to the substrate, and hole tunneling has a much lower probability. If one assumes that the total active gate area per chip is of the order of 0.1 cm^2, the maximum tolerable gate leakage current will be of the order of 10 A/cm^2. This sets a lower limit of 1.0–1.5 nm for the gate-oxide thickness. Dynamic memory devices have a more stringent leakage requirement and therefore must impose a higher limit on gate-oxide thickness [6].

Another issue with the thin gate oxide is the loss of inversion charge and therefore transconductance due to inversion-layer quantization and polysilicon-gate depletion effects [2]. Quantum mechanics dictates that the density of inversion electrons peaks at approximately 1 nm below the silicon surface, which effectively reduces the gate capacitance and therefore the inversion charge to those of an equivalent oxide ~0.4 nm thicker than the physical oxide [4]. Similarly, depletion effects occur in polysilicon in the form of a thin space-charge layer near the oxide interface which acts to reduce the gate capacitance and inversion-charge density for a given gate drive. The percentage of gate-capacitance attenuation becomes more significant as the oxide thickness is scaled down. For a polysilicon doping of 10^{20} cm^{-3}, a 2-nm oxide loses about 20% of the inversion charge at 1.5-V gate voltage because of the combined effects of polysilicon gate depletion and inversion-layer quantization [2]. Taking these two effects into account, the scaling limit of the electrical oxide thickness (t_{inv}), i.e., the effective oxide thickness for inversion charge calculations, is likely to be 1.5–2.0 nm.

3. Design considerations near the scaling limit

From the above discussions, CMOS design space is severely constrained by voltage and oxide limits below 100-nm dimensions. In this section, we consider a 2D scale length model and present an optimized 2D nonuniform doping profile design that can extend CMOS scaling to a minimum channel length of 20 nm.

2D scale length theory

Figure 4 shows the essential 2D aspects of a short-channel MOSFET [7]. A key parameter is the gate depletion width, W_d, within which the mobile carriers (holes in the case of n-MOSFETs) are swept away by the applied gate field. The gate depletion width reaches a maximum, W_{dm}, at the onset of strong inversion (threshold voltage) when the surface potential (ψ_s) or band bending is such that the electron concentration at the surface equals the hole concentration in the bulk substrate. This is the standard $\psi_s = 2\psi_B$ condition, with $\psi_B = (kT/q)\ln(N_a/n_i)$, where N_a is the substrate doping concentration and n_i the intrinsic carrier concentration of silicon. For uniformly doped cases,

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si}kT\ln(N_a/n_i)}{q^2N_a}}. \quad (3)$$

A rectangle is formed by the boundary of the gate depletion region, the gate electrode, and the source and drain regions, as depicted in Figure 4 [7]. Two-dimensional effects can be characterized by the aspect ratio of this rectangle. When the horizontal dimension, i.e., the channel length, is at least twice as long as the vertical dimension, the device behaves like a long-channel MOSFET, with its threshold voltage insensitive to channel length and drain bias. For channel lengths shorter than that, the 2D effect becomes significant, and the minimum surface potential (ψ_s) which determines the threshold

voltage is increasingly controlled more by the drain than by the gate.

The rectangular box consists of a silicon region of thickness W_{dm} and an oxide region of thickness t_{ox}. At the interface, the vertical fields ($\mathcal{E}_x$) obey the boundary condition, $\varepsilon_{si}\mathcal{E}_{x,si} = \varepsilon_{ox}\mathcal{E}_{x,ox}$, where ε_{si}, ε_{ox} are the permittivities of silicon and oxide, respectively. For lateral fields ($\mathcal{E}_y$) tangential to the interface, the boundary condition is $\mathcal{E}_{y,si} = \mathcal{E}_{y,ox}$, independent of the dielectric constants. By properly matching the boundary conditions of both components of electric fields at the silicon–insulator interface, one can derive a scale length λ that is a solution to the following equation [8]:

$$\varepsilon_{si}\tan(\pi t_{ox}/\lambda) + \varepsilon_{ox}\tan(\pi W_{dm}/\lambda) = 0. \quad (4)$$

The scale length λ also goes into the length-dependent term of the maximum potential barrier in a MOSFET of channel length L, i.e., $\Delta\psi_s(\text{SCE}) \propto \exp(-\pi L/2\lambda)$ [7]. The ratio L/λ is a good measure of the strength of the 2D effect. For the short-channel V_t roll-off and the drain-induced barrier lowering (DIBL) to be acceptable, the above exponential factor must be much less than 1. This means that the minimum useful channel length is about 1.5–2.0 times λ [4]. Note that the above scale-length equation is valid for the high-k gate insulator as well. One simply replaces ε_{ox} and t_{ox} with ε_i and t_i, where ε_i is the permittivity of that insulator and t_i its thickness.

The lowest-order solution to the above equation is plotted in **Figure 5** in the form of constant-λ contours in a t_{ox}–W_{dm} design plane. In addition to the 2D scale-length requirement, the ratio between t_{ox} and W_{dm} must also be kept small in order for the inverse (log) subthreshold current slope [Equation (1)] [4],

$$S = m(\ln 10)\frac{kT}{q} = \left(1 + \frac{\varepsilon_{si}t_{ox}}{\varepsilon_{ox}W_{dm}}\right)(\ln 10)\frac{kT}{q}, \quad (5)$$

to be close to the ideal $(\ln 10)kT/q$ value, or 60 mV/decade. Here m is usually referred to as the ideality factor which measures the gate-voltage swing required per unit of change in the electron potential (or band bending) at the silicon surface. A reasonable upper limit is $t_{ox}/W_{dm} = 0.1$, or $m = 1.3$, as indicated by the dotted curve in Figure 5. This gives a long-channel inverse subthreshold slope of ≈80 mV/decade. The intercepts of the dotted curve with the constant-λ contours lie in a region where the vertical fields dominate, and $\lambda \approx W_{dm} + (\varepsilon_{si}/\varepsilon_{ox})t_{ox}$, obtained by replacing the oxide region with an equivalent silicon region of thickness $(\varepsilon_{si}/\varepsilon_{ox})t_{ox}$ [7]. The design points, or the intercepts, can then be solved as $t_{ox} = (1 - 1/m)(\varepsilon_{ox}/\varepsilon_{si})\lambda$. This means that for $L_{min} \approx (1.5\text{–}2.0)\lambda$ and $m \approx 1.3$, the oxide thickness required is $t_{ox} \approx L_{min}/20$ to $L_{min}/25$ [4].

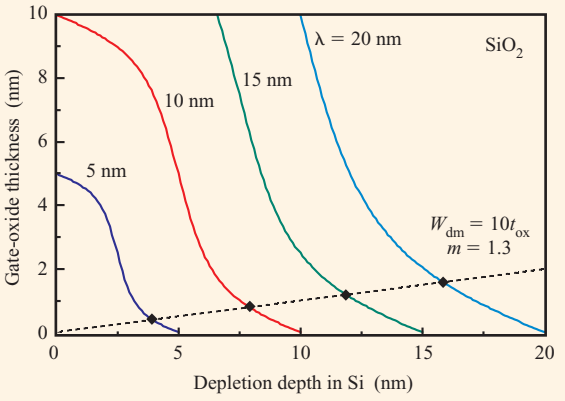

Figure 5

Constant scale length λ contours (solid lines) in a t_{ox}–W_{dm} design plane, assuming SiO_2 or $\varepsilon_{Si}/\varepsilon_{ox} = 3$. The dotted line marks the boundary on which the ideality factor m equals 1.3. The intercepts represent design points which satisfy both the scale length and the subthreshold slope requirements.

25-nm CMOS design

The discussions in Section 2 make it clear that CMOS devices below 0.1 μm will have increasing leakage currents, declining performance gains, and higher chip power. Nevertheless, with proper control of the doping profile, the limit of CMOS scaling can be extended to 20-nm channel length without strict scaling of oxide thickness and power-supply voltage (Figure 1).

An optimum design for 20-nm MOSFET calls for a vertically and laterally nonuniform doping profile, the *superhalo* [9], to control the short-channel effect. **Figure 6** shows such a doping profile, along with simulated potential contours for a 25-nm MOSFET [9]. Halo doping, or nonuniform channel profile in the lateral direction, can be realized by angled ion implantation self-aligned to the gate, with a very restricted amount of diffusion. The highly nonuniform profile sets up a higher effective doping concentration toward shorter devices, which counteracts short-channel effects. With this design, the simulated off-currents are insensitive to channel-length variations between 20 and 30 nm, as shown in **Figure 7**. In terms of the threshold-voltage sensitivity to channel-length variations, the superhalo profile extends the scaling limit by a factor of nearly 2. For example, considering a criterion of $\Delta V_t \leq 50$ mV for $\Delta L/L = 30\%$, the superhalo profile in Figure 6 ($\lambda \approx 17$ nm) allows for $L_{min} = 20$ nm $\approx 1.2\lambda$, while a non-halo MOSFET can be scaled to only $L_{min} \approx 2\lambda$ against the same criterion. From another perspective, because of the flat V_t dependence on channel length, the superhalo profile permits a nominal device to

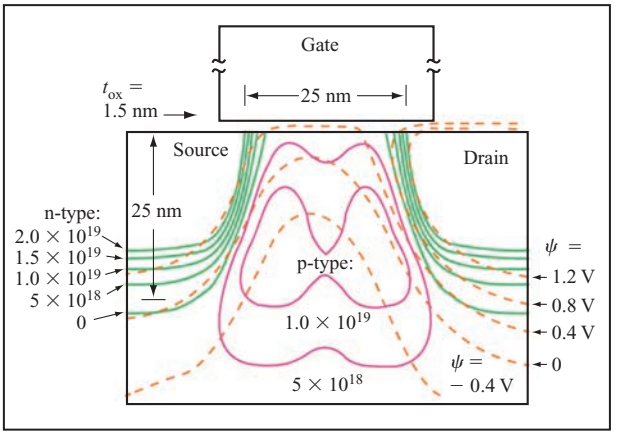

Figure 6

Source, drain, and superhalo doping contours in a 25-nm n-MOSFET design. The channel length is defined by the points at which the source–drain doping concentration falls to 2×10^{19} cm^{-3}. The dashed curves show the potential contours for zero gate voltage and a drain bias of 1.0 V. $\psi = 0$ refers to the midgap energy level of the substrate. Reprinted with permission from [9]; ©1998 IEEE.

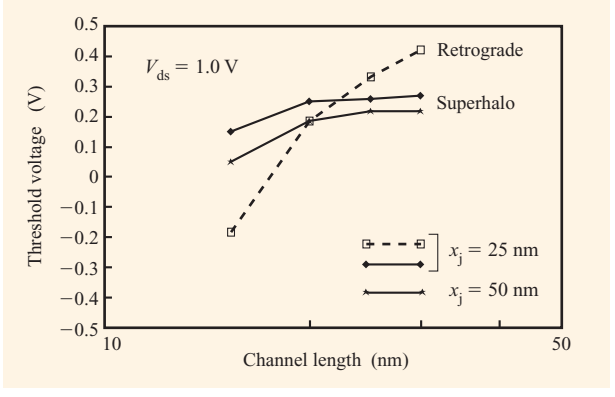

Figure 7

Short-channel threshold roll-off for superhalo and retrograde (non-halo) doping profiles. For the superhalo case, two junction depths, x_j = 25 nm (that of Figure 6) and x_j = 50 nm, were studied and compared. Threshold voltage is defined as the gate voltage where $I_{ds} = 1\ \mu$A/μm. Reprinted with permission from [9]; ©1998 IEEE.

operate at a lower threshold voltage, thereby gaining significant performance benefit [9].

The designed profile in Figure 6 is forgiving with respect to the junction depth. Figure 7 shows that the V_t roll-off is rather insensitive to the vertical junction depth, with only a slight change when the junction depth is doubled from 25 nm to 50 nm for the same halo profile. This points to a way out of the high-resistance problem associated with very shallow extensions [10]. The lateral source–drain gradient, however, is much more critical. The short-channel effect degrades rapidly once the profile is more graded than 4–5 nm/decade. This is because the channel length is largely determined by the points of current injection from the surface layer into the bulk, which takes place at a source–drain doping concentration of about 2×10^{19} cm^{-3} [11]. Any source–drain doping that extends beyond this point into the channel tends to compensate or counterdope the channel region and aggravate the short-channel effect. The abruptness requirements of both the source–drain and halo doping profiles dictate absolutely minimum thermal cycles after the implants. Note that a raised source–drain structure may help in making contacts, but does not by itself satisfy the abruptness requirement discussed here.

One of the concerns with the high p-type doping level and narrow depletion regions in Figure 6 is the band-to-band tunneling through the high-field region between the p-halo and the reverse-biased drain. For a drain voltage of 1 V, the highest field at the heavily doped halo region is estimated to be 1.7 MV/cm. According to the published data, the band-to-band tunneling current density of the drain-to-substrate junction for this field magnitude is of the order of 1 A/cm^2 [9]. This should not constitute a major component of the device leakage current, given the narrow width of the high-field region (Figure 6).

The threshold design in Figure 7 assumes dual n^+/p^+ Si work-function gates for n-MOS/p-MOS, respectively. A midgap-work-function metal gate would clearly result in threshold voltage magnitudes far too high for both devices [4]. Unless dual metal gates having work functions comparable to those of n^+/p^+ poly can be developed, one must deal with the effect of poly depletion on CMOS performance. Since the capacitance of the poly depletion layer is not a constant, but depends on both the gate voltage and the quasi-Fermi potential along the channel, treating it as an equivalent-oxide layer will substantially overestimate its effect [4]. Comparisons between a poly-gated and a metal-gated device show that while typical C–V data of 1.5-nm oxides exhibit about 40% less capacitance at inversion than that of the physical oxide, the currents are degraded by only 10–20% [9]. One factor is that part of the capacitance loss comes from quantum-mechanical effect in the inversion layer, which is present regardless of the gate material [2]. Another factor is that while the inversion charge density is higher in metal-gate devices, carrier mobilities are lower because of the higher vertical field (like having a thinner oxide [4]). The effect of poly-gate depletion on CMOS circuit delay is even less. The intrinsic, unloaded inverter delay is only slightly degraded (≈5%) [9] because although poly depletion causes a loss in the drive current, it also decreases the

charge needed for the next stage. These two effects tend to cancel each other. For the heavily loaded case in which the devices drive a large fixed capacitance, the delay degradation approaches those of the on-currents. However, this can be compensated to some extent by using wider devices.

Three-dimensional statistical simulations have been carried out on the effects of dopant fluctuations on threshold voltage in 25-nm CMOS. For the doping profile in Figure 6, dopant number fluctuations cause a $10/\sqrt{W}$ mV-μm$^{1/2}$ (1σ) uncertainty in the threshold voltage [12], where W is the device width. Compared with threshold tolerances from short-channel effects, which do not depend on device width, this number is small for relatively wide (~1-μm) devices in logic circuits. However, for minimum-width (~0.1-μm) devices in SRAM cells, the 6σ (needed for ~100-Mb arrays) threshold variation due to dopant number fluctuations approaches 200 mV. This clearly must be taken into consideration during design in order to obtain reasonable yields.

To evaluate the potential on-state performance of 25-nm CMOS, detailed Monte Carlo simulations were performed using the simulator DAMOCLES [13]. Both n- and p-channel MOSFETs have been simulated, yielding low-output-conductance, high-performance I–V characteristics for both device types. The n-FET transconductance exceeds 1500 mS/mm, with an estimated f_T higher than 250 GHz [9]. Transient Monte Carlo simulations were also done for a three-stage chain of 25-nm CMOS inverters, giving a delay time of 4–4.5 ps for V_{dd} = 1.0 V and I_{off} = 1 nA/μm. Lower threshold voltages (higher off-currents) would result in somewhat shorter delays.

4. Extending CMOS scaling to 10 nm

It was discussed in previous sections that the CMOS performance trend will slow below the 100-nm channel length because of fundamental factors of oxide tunneling and voltage nonscaling. One option that offers opportunities for further gains in performance-driven systems is cooled CMOS. The benefits are derived primarily from two aspects of MOSFET characteristics at low temperature: higher carrier mobilities and steeper subthreshold slope. Electron mobility improves by a factor of 2 to 5 from 300 K to 77 K, depending on the magnitude of the vertical field [14]. Similarly, hole mobility also improves by a factor of 1.7 to 4 for the same temperature range. In addition, MOSFET subthreshold slope steepens by a factor inversely proportional to the absolute temperature, making it much easier to turn off a device at low temperature than at room temperature (**Figure 8**) [15]. This allows the threshold voltage V_t, and therefore the power-supply voltage V_{dd}, to scale down further below their permissible values at room

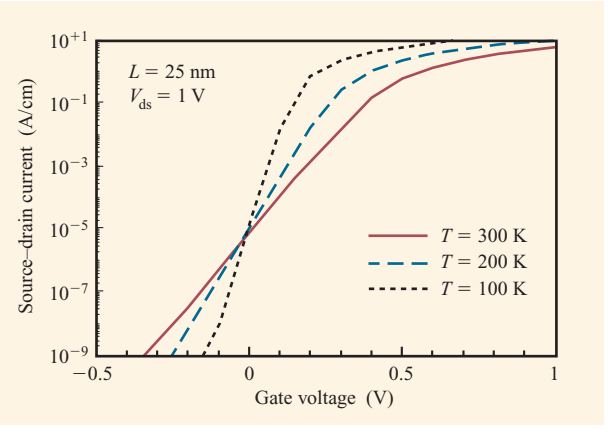

Figure 8

Simulated subthreshold currents of 25-nm MOSFETs at three different temperatures. In each case, V_t is adjusted by doping to maintain the same I_{off} at V_g = 0.

temperature. However, to operate at a low nominal V_t requires that both the worse-case minimum V_t and the V_t tolerances such as those due to short-channel effects be reduced. While low-temperature operation allows for a lower minimum V_t, it offers no relief for short-channel V_t tolerances, which are controlled by electrostatics independent of temperature. It is therefore essential in low-temperature CMOS to use optimized doping profiles, i.e., superhalo, to tighten the threshold-voltage tolerances [9].

Because more band bending is needed to reach the $2\psi_B$ threshold condition as the temperature decreases, the threshold voltage of a given CMOS hardware increases at lower temperatures [4]. Thus, for CMOS devices fabricated with acceptable off-currents at, e.g., 100°C, the threshold voltage would be too high at low temperatures. To gain the most performance from low-temperature CMOS, one should turn the threshold voltage trend around and take advantage of the steeper subthreshold slope. This is illustrated by the *same off-current* design in Figure 8, in which the threshold voltages are adjusted to lower values as the temperature decreases, such that the off-current is maintained at the same level as the product specification at 100°C [16]. This can be accomplished to some extent using a retrograde channel profile without degrading the short-channel effect.

If the temperature is low enough, some of the subthreshold slope steepness can be traded off for a shorter scale length λ and therefore better control of the short-channel effect. It works because with the smaller kT/q, one can allow the ideality factor m in Equation (5) to be substantially larger than 1 while still having a respectable slope S. So, for a given t_{ox}, W_{dm} can take on

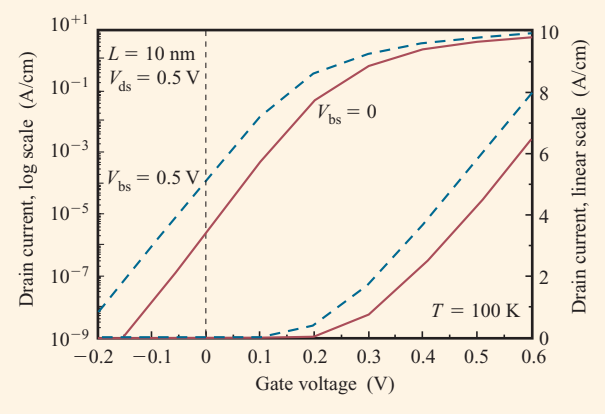

Figure 9

Simulated drain currents of a 10-nm MOSFET (t_{ox} = 1.5 nm) at 100 K with (dashed curve) and without (solid curve) forward substrate bias. Same currents are plotted in both the log scale (left) for reading of I_{off} and the linear scale (right) for reading of I_{on}.

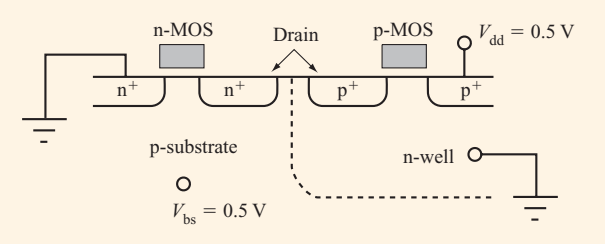

Figure 10

Schematic bias diagram of low-temperature 10-nm CMOS with V_{dd} = 0.5 V and a forward body bias of V_{bs} = V_{dd} = 0.5 V for both n- and p-MOSFETs. In CMOS inverters, the gates are tied together as the input and the drains are tied together as the output.

values smaller than that bounded by the dotted line in Figure 5. This opens up the possibility of extending CMOS scaling to 10-nm channel length. **Figure 9** shows a simulated design example of an n-MOSFET with L = 10 nm operated at V_{dd} = 0.5 V and 100 K [17]. Here the inverse subthreshold slope, with some short-channel degradation, is about 45 mV/decade, and the ideality factor, m, is in the range of 1.5 to 2.0. A nonscaled oxide thickness of 1.5 nm (including poly depletion effects, if any) is assumed.

Two issues must be addressed in this scenario: band-to-band tunneling and the high magnitude of threshold voltage. Both can be relieved by applying a forward bias to the body, as depicted in Figure 9 and **Figure 10**. For an ideal super-steep retrograde profile with n^+ poly gate on n-MOSFET, the long-channel threshold voltage is [4]

$$V_t = V_{fb} + 2m\psi_B - (m-1)V_{bs} \approx (m-1)(E_g - V_{bs}). \quad (6)$$

Here V_{fb} is the flatband voltage and V_{bs} is the bias voltage of the p-type body. In the last part of Equation (6), $2\psi_B$ is approximated by the silicon bandgap E_g, and V_{fb} by $-E_g$. So an m value of, e.g., 1.5 to 2 would increase V_t, while a forward bias of, e.g., V_{bs} = 0.5 V would lower V_t. Other more elaborate means to reduce V_t include counterdoping of the channel and/or using gate work functions outside of those of n^+/p^+ poly. In the scheme shown in Figure 10, in which V_{bs} = V_{dd} = 0.5 V, both drain junctions have zero bias when the corresponding device is in the off state, thus circumventing the band-to-band tunneling problem with such narrow depletion widths. Although the substrate- or well-to-source junctions (and to drain when the device is on) are forward-biased to V_{dd} [17], the forward-bias currents should be at a negligible level, since V_{dd} = 0.5 V is far below the turn-on voltage (~0.9 V) of a p–n junction at low temperatures.

For the case shown in Figure 9, the low V_t of ≈0.2 V is obtained with a combination of forward body bias and some short-channel V_t roll-off at 10 nm channel length. While the off-currents appear acceptable and the on-currents respectable for a V_{dd} of 0.5 V, the short-channel effect is rather poor, since t_{ox} has not been scaled and since low-temperature operation provides no relief from 2D effects. This should clearly be considered as the worst-case L_{min}. Circuit noise margins may become a significant issue as drain-induced barrier lowering and output conductance deteriorate. Furthermore, source-to-drain tunneling [18] through the potential barrier in the inset of Figure 2, rather than thermal injection over the barrier, may become the dominant leakage mechanism at such low temperatures.

5. Conclusion

In conclusion, CMOS scaling below 100-nm channel length faces several fundamental limiting factors stemming from electron thermal energy and quantum-mechanical tunneling. Many of the potential barriers in a MOSFET that kept the standby leakage low are losing their effectiveness when scaled to lower barrier heights or thinner widths. Inevitably, both the standby power and the active power of a high-performance processor will rise. As a tradeoff, the performance gained from scaling will slow. Nevertheless, by using properly optimized doping profiles and pushing the silicon depletion width to the tunneling limit, it is likely that mainstream CMOS scaling will be extended to 20-nm channel length with nonscaled gate oxides and voltage levels. Beyond that, cooling to low temperature might provide the additional design space needed to extend CMOS devices to 10 nm for server applications.

Acknowledgment

The author would like to thank David Frank, Bob Dennard, and Paul Solomon for many stimulating discussions.

References

1. R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits* **SC-9,** 256 (1974).
2. Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H.-J. C. Wann, S. J. Wind, and H.-S. Wong, "CMOS Scaling into the Nanometer Regime," *Proc. IEEE* **85,** 486–504 (1997).
3. 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,** 209–211 (1997).
4. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, New York, 1998.
5. Y. Taur, Y.-J. Mii, D. J. Frank, H.-S. Wong, D. A. Buchanan, S. J. Wind, S. A. Rishton, G. A. Sai-Halasz, and E. J. Nowak, "CMOS Scaling into the 21st Century: 0.1 μm and Beyond," *IBM J. Res. & Dev.* **39,** 245–260 (1995).
6. 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).
7. T. N. Nguyen, "Small-Geometry MOS Transistors: Physics and Modeling of Surface- and Buried-Channel MOSFETs," Ph.D. Thesis, Stanford University, California, 1984.
8. D. J. Frank, Y. Taur, and H.-S. Wong, "Generalized Scale Length for Two-Dimensional Effects in MOSFET's," *IEEE Electron Device Lett.* **19,** 385 (1998).
9. Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS Design Considerations," *IEDM Tech. Digest*, p. 789 (1998).
10. S. Thompson, P. Packan, T. Ghani, M. Stettler, M. Alavi, I. Post, S. Tyagi, S. Ahmed, S. Yang, and M. Bohr, "Source–Drain Extension Scaling for 0.1 μm and Below Channel Length MOSFETs," *Symposium on VLSI Technology, Digest of Technical Papers*, 1998, p. 132.
11. Y. Taur, Y. J. Mii, R. Logan, and H. S. Wong, "On Effective Channel Length in 0.1 μm MOSFETs," *IEEE Electron Device Lett.* **16,** 136 (1995).
12. D. J. Frank, Y. Taur, M. Ieong, and H.-S. Wong, "Monte Carlo Modeling of Threshold Variation Due to Dopant Fluctuations," *Symposium on VLSI Technology, Digest of Technical Papers*, 1999, p. 169.
13. S. E. Laux, M. V. Fischetti, and D. J. Frank, "Monte Carlo Analysis of Semiconductor Devices: The DAMOCLES Program," *IBM J. Res. & Dev.* **34,** 466–494 (1990).
14. S. Takagi, M. Iwase, and A. Toriumi, "On Universality of Inversion-Layer Mobility in n- and p-Channel MOSFETs," *IEDM Tech. Digest*, pp. 398–401 (1998).
15. F. H. Gaensslen, V. L. Rideout, E. J. Walker, and J. J. Walker, "Very Small MOSFETs for Low Temperature Operation," *IEEE Trans. Electron Devices* **ED-24,** 218 (1977).
16. Y. Taur and E. J. Nowak, "CMOS Devices Below 0.1 μm: How High Will Performance Go?" *IEDM Tech. Digest*, p. 215 (1997).
17. Y. Taur, "CMOS Scaling Beyond 0.1 μm: How Far Can It Go?" *Digest of Technical Papers, International Symposium on VLSI Technology, Systems, and Applications*, Taipei, Taiwan, June 1999, pp. 6–9.
18. Y. Naveh and K. K. Likharev, "Modeling of 10-nm-Scale Ballistic MOSFETs," *IEEE Electron Device Lett.* **21,** 242 (2000).

Received June 22, 2001; accepted for publication December 5, 2001

Yuan Taur *Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, California 92093 (taur@ece.ucsd.edu).* Dr. Taur received the B.S. degree in physics from the National Taiwan University, Taipei, Taiwan, in 1967 and the Ph.D. degree in physics from the University of California, Berkeley, in 1974. From 1975 to 1979, he worked at NASA, Goddard Institute for Space Studies, New York, on low-noise Josephson junction mixers for millimeter-wave detection. From 1979 to 1981, he worked at Rockwell International Science Center, Thousand Oaks, California, on II–VI semiconductor devices for infrared sensor applications. From 1981 to 2001, Dr. Taur was with the Silicon Technology Department of the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, where he was Manager of Exploratory Devices and Processes. Areas in which he has worked and published include latchup-free 1-μm CMOS, self-aligned $TiSi_2$, 0.5-μm CMOS and BiCMOS, shallow-trench isolation, 0.25-μm CMOS with n^+/p^+ polysilicon gates, SOI, low-temperature CMOS, and 0.1-μm CMOS. Since October 2001, he has been a professor in the Department of Electrical and Computer Engineering of the University of California at San Diego. Dr. Taur was elected a Fellow of the IEEE in 1998. He is currently the Editor-in-Chief of the *IEEE Electron Device Letters*. He has served on the technical program committees and as a panelist for the Device Research Conference and the International Electron Devices Meeting, and as Program Chairman for the Symposium on VLSI Technology. Dr. Taur has authored or co-authored more than 120 technical papers; he holds 11 U.S. patents. He is profiled in the 2001 edition of *Who's Who in the World*. Dr. Taur received four IBM Outstanding Technical Achievement Awards and six IBM Invention Achievement Awards during his IBM career. He co-authored the book, *Fundamentals of Modern VLSI Devices*, published by Cambridge University Press in 1998.