CMOS Scaling for High Performance and Low Power—The Next Ten Years

BIJAN DAVARI, SENIOR MEMBER, IEEE, ROBERT H. DENNARD, FELLOW, IEEE, AND GHAVAM G. SHAHIDI

Invited Paper

A guideline for scaling of CMOS technology for logic applications such as microprocessors is presented covering the next ten years, assuming that the lithography and base process development driven by DRAM continues on the same three-year cycle as in the past. This paper emphasizes the importance of optimizing the choice of power-supply voltage. Two CMOS device and voltage scaling scenarios are described, one optimized for highest speed and the other trading off speed improvement for much lower power. It is shown that the low power scenario is quite close to the original constant electric-field scaling theory. CMOS technologies ranging from 0.25 μm channel length at 2.5 V down to sub-0.1 μm at 1 V are presented and power density is compared for the two scenarios. Scaling of the threshold voltage along with the power-supply voltage will lead to a substantial rise in standby power compared to active power, and some tradeoffs of performance and/or changes in design methods must be made. Key technology elements and their impact on scaling are discussed.

It is shown that a speed improvement of about 7× and over two orders of magnitude improvement in power-delay product (mW/MIPS) are expected by scaling of bulk CMOS down to the sub-0.1 μm regime as compared with today's high performance 0.6 μm devices at 5 V. However, the power density rises by a factor of 4× for the high-speed scenario. The status of the silicon-on-insulator (SOI) approach to scaled CMOS is also reviewed, showing the potential for about 3× savings in power compared to the bulk case at the same speed.

I. INTRODUCTION

The growth of microelectronics during the last 25 years, from the first successful large-scale integration (LSI) of microprocessor and memory chips to the present ultra large-scale integration (ULSI), has been truly spectacular. The key to this growth has been the drive to much smaller dimensions using the principles of scaling introduced in the early 1970's [1] which we briefly review here. The basic idea of scaling, shown in Fig. 1, is to reduce the dimensions of the MOS transistors and the wires connecting them in integrated circuits. Thus the arrangement on the right is scaled down in size from that on the left by reducing all dimensions by a factor α. The MOS transistor works on the principle of modifying the electric field in the silicon substrate underneath the gate in such a way as to control the flow of current between the source and drain electrodes. Scaling achieves the same electric-field patterns in the smaller transistor by reducing the applied voltage along with all the key dimensions, including the thickness, t_{ox} of the insulating oxide layer between the gate and the silicon substrate. Within the silicon substrate, the electric field patterns are preserved by increasing the impurity doping concentration of the smaller device. Taken along with the reduced applied voltage, this reduces the size of the depletion regions, identified by x_d in Fig. 1, underneath all three transistor electrodes. In general, these depletion regions must be kept separated so that the transistor can be turned off properly by the control gate [2]. The scaled-down depletion regions in the transistor on the right of Fig. 1 allows the separation L between source and drain to be reduced along with the other physical dimensions. In this simple constant-electric-field transformation, the dimension, voltage, and doping are all modified by a common factor α, as noted in the figure.

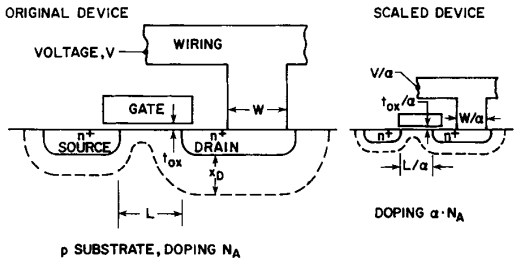


Fig. 1. Principles of constant-electric-field scaling for MOS transistors and integrated circuits.

This constant-electric-field scaling gives three important results. First, the density improves by a factor α^2 due to

Manuscript received October 13, 1994; revised December 8, 1994.

The authors are with the IBM Semiconductor Research and Development Center (SRDC), T. J. Watson Research Center, Yorktown Heights, NY 10598 USA.

IEEE Log Number 9408754.

Table 1 Generalized Scaling Relationships

Physical Parameters	Constant-Electric Field Scaling Factor	Generalized Scaling Factor
Linear Dimensions	$1/\alpha$	$1/\alpha$
Electric-Field Intensity	1	ϵ
Voltage (Potential)	$1/\alpha$	ϵ/α
Impurity Concentration	α	$\epsilon\,\alpha$

the smaller wiring and device dimensions. Next, the speed, which is related to g_m/C, improves by a factor α because the capacitance (C) of the shorter wires and smaller devices is reduced by d while the transconductance (g_m) of the devices (scaled in both length and width) remains about the same. Finally, the power dissipation per circuit is reduced by a factor α^2 because of the reduced voltage and current in each device, with the important result that the power density is constant. Thus the increased number of circuits in a given chip area can be accommodated with no increase in the total power dissipation.

Although the original concept of constant-electric-field (CE) scaling is useful and valid, the idea of reducing voltage in proportion to reduced dimensions has not been popular because of reluctance to depart from standardized voltage levels. Also, scaling the threshold voltage of the devices down along with the applied voltage increases the standby leakage current, which limits how far it is practical to scale the power-supply voltage [1]. Therefore, it was useful to broaden the concept of scaling to the more generalized form shown in Table 1, where the electric-field patterns within a scaled device are still preserved, but the intensity of the electric field can be changed everywhere within the device by a multiplicative factor ϵ [3]. Thus the applied voltage, which is given by ϵ/α, can be scaled less rapidly by allowing ϵ to increase. The electric-field patterns within the device are maintained by increasing the doping impurity concentration by a factor ϵ, which preserves the size of the depletion regions, x_d, defined in Fig. 1.

Obviously, there are practical limits to generalized scaling. Increased electric field $(\epsilon > 1)$ is limited by reliability effects during long-term use such as device degradation resulting from hot-carrier mechanisms or gate-insulator failure. Ideally, the current in a scaled device or circuit increases by a factor ϵ^2 and the speed by ϵ, up to the point where it is limited by carrier velocity saturation effects and by the increased series resistance of graded junctions needed for hot-carrier reliability at the higher voltage levels. Because of such reliability considerations, the speed can actually decrease as the electric-field parameter ϵ is increased beyond a practical value, as will be discussed in the next section [4]. Another very significant limit to ϵ is the power dissipation, which increases by ϵ^2 when the speed is constant (as given by the familiar power calculation CV^2f).

The application of scaling to CMOS technology up to now is illustrated in Fig. 2, which shows the typical loaded NAND delay versus L for three logic products. The gate oxide thickness is scaled nearly linearly with channel length as shown, while the voltage levels (including the threshold voltage) are reduced by approximately the square root of L. It is seen that even with the much higher electric

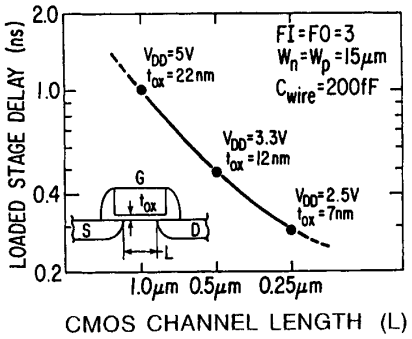


Fig. 2. Scaling of high-performance CMOS technology.

field in the shorter devices, the delay improves only about linearly with channel length due to the velocity saturation and resistance effects discussed above. While the measured performance in Fig. 2 used test circuits with a constant wiring capacitance and device widths, the quoted values are only typical of the $L = 1$ μm generation and will both be scaled in proportion to the lithography dimensions in the real chips in the scaled technologies. Therefore, since both the wiring capacitance and device widths shrink simultaneously in scaled technologies, the delay numbers shown in Fig. 2 remain an accurate measure of the technology's performance.

Although the scaling rules in Fig. 1 show the device and wiring dimensions being scaled by the same factor, they can actually be scaled by different factors. We call this "selective scaling," where the device channel length and t_{ox} are scaled by a factor α_d while the channel width and wiring width are scaled by another factor α_w. The speed is increased in correspondence with the device scaling factor, α_d, while the density improves by α_w^2 and the power per circuit decreases by a factor $\alpha_d\alpha_w/\epsilon^2$ assuming the voltage levels are scaled by ϵ/α_d. For the CMOS logic generations of Fig. 2 the typical minimum lithography dimensions used for interconnections are 1.25, 0.8, and 0.5 μm, respectively.

In the future evolution of CMOS the scaling of voltage levels will become a crucial issue. A new paradigm in the information industry is taking shape which will allow and demand much more frequent voltage reductions down to as low as 1 V or less. The main forces behind this shift are the ability to produce complex, high-performance systems on a single chip and the projected explosion in demand for portable and wireless systems with very low power budgets. Both of these will allow an unprecedented degree of freedom in choosing the supply voltages for the IC's due to self-containment of massive information processing capability. In addition, various memory and ASIC's will also embrace lower supply voltages to maintain manageable power densities. However, the increase in standby leakage current that will result from further scaling of threshold voltage poses a severe challenge, particularly for battery-powered applications.

In the following sections two different scaling scenarios for CMOS devices and logic circuits are discussed as dif-

ferent priorities in the "speed/power/reliability/density/cost" design space. The average longitudinal electric field along the device channel is used as a parameter in projecting the evolution of device miniaturization for a high-performance scenario and for a low-power scenario and in their comparison with the constant electric field (CE) scaling. Next, we discuss the key technology elements which enable the fabrication of the scaled technologies, followed by a discussion of scaled CMOS on silicon on insulator (SOI) which offers significant performance and power improvements over CMOS on bulk.

II. Future CMOS Device Scaling

As miniaturization of CMOS devices progresses without reducing the power-supply voltage as much as proposed by the CE scaling, the electric field increases substantially. This section begins with a discussion of how high the electric field can be without impacting the long term device reliability, while at the same time achieving high performance (speed) and reasonable power. The tradeoff between speed and power dissipation determines how aggressively the supply voltage should be scaled down, while the reliability constraints set an upper limit on the useful power-supply voltage. Scenarios for design point optimization are given for both high-performance and very low-power applications. Then the speed/standby-current tradeoff is addressed, dealing with the issue of nonscalability of the threshold voltage. The impact of the two scaling scenarios on power density is evaluated. Finally, a summary set of CMOS logic scaling guidelines extending for the next ten years is given and the important projected results in terms of performance, power, and density are highlighted.

A. Performance/Voltage/Reliability Tradeoff

The effect of the channel hot-carrier (CHC) limits on the choice of the optimum power-supply voltage for 0.25 μm CMOS is shown in Fig. 3. For this figure, an empirical relationship between CHC limited voltage and device series resistance, added by drain engineering, was developed from measurements on a variety of junction technologies [4]. Using this relationship with a model of the intrinsic device behavior, the circuit performance as a function of the power-supply voltage and CHC margin can be determined. A certain amount of CHC margin, about 0.5 V, depending on the conditions, is needed to allow defect screening with accelerated applied voltage and temperature (burn-in) without damaging the devices. For a constant CHC reliability margin, an optimum supply voltage exists, above which the CMOS performance actually degrades due to the excessive added device series resistance which is needed to support the high voltage operation. At power-supply voltages higher than 2.5 V, in order to limit the long-term reliability impact, drain engineering, e.g., lightly doped drain (LDD), is needed for this 0.25 μm channel length. The LDD structures allow reliable operation at higher electric fields [5], but can result in the reduction of the device performance due to increased device series

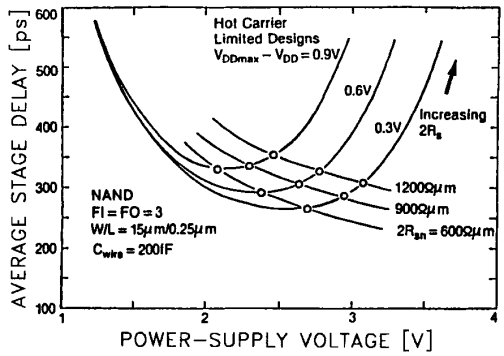


Fig. 3. Performance versus reliability tradeoff for $L = 0.25$ μm CMOS with reliability margin (V_{DD} max $-V_{DD}$) as a parameter.

resistance. Therefore the choice of the supply voltage and the appropriate device profiles represent optimization of a tradeoff between performance and reliability. We should note that the points on the constant reliability margin contours in Fig. 3 are not from the same technology, rather they represent devices with varying series resistance with similar channel lengths. Also, an oxide thickness of 7 nm is used, which gives adequate defect density and reliability at the 2.5 V design point [31]. The reduction of the oxide thickness relative to the previous 3.3 V technologies (Fig. 2) is possible due to the lower power-supply voltage.

As we scale the CMOS technology beyond 0.25 μm, in order to obtain a significant performance improvement of about 1.5× per generation, a reduced source/drain (S/D) spreading resistance is needed which requires S/D junctions with more abrupt profiles [7]. Therefore the drain electric field is increased which results in a lower optimum power-supply voltage for a given reliability margin.

In summary, the CHC degradation dictates an optimum power-supply voltage for high-speed logic technologies. Above this voltage the performance actually degrades due to excessive S/D resistance which is needed to maintain the reliability margin. This optimum voltage is 2.5 V for 0.25 μm devices and it decreases for smaller channel lengths. However, Fig. 3 also clearly shows that even lower voltage can be used with a modest decrease in speed, providing the possibility of significantly lower power consumption.

B. High-Performance and Low-Power Voltage Scaling Scenarios

Taking into account the considerations discussed above and other factors, two scenarios for scaling down power-supply voltage in future scaled CMOS logic technologies are proposed. It is instructive to describe these scenarios in the context of generalized scaling as shown in Fig. 4, which plots the electric field as a function of channel length where the measure of electric field is taken to be the applied power-supply voltage divided by the channel length of the CMOS devices. The three curves in Fig. 4 correspond to the following:

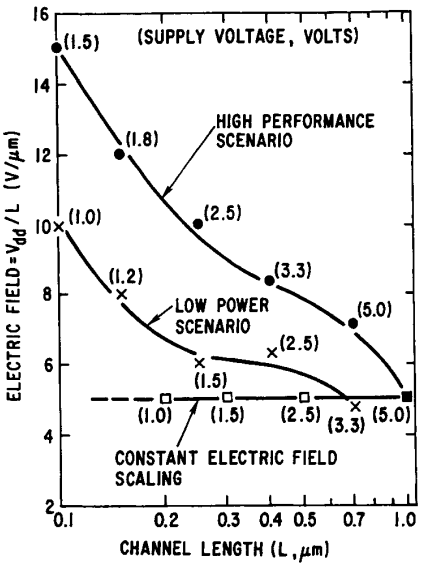


Fig. 4. A measure of the electric field, V_{DD}/L, as a function of channel length for various scaling scenarios.

The first curve gives the high performance scenario, where the power-supply voltage at each channel length is optimized for maximum speed while maintaining adequate long-term device reliability. Other factors in this optimization are the lithography tolerance at the gate level, which is assumed to be 20% of the nominal mask gate length (L_{mask}), and the threshold voltage, which must be high enough for acceptable leakage current (I_{off}) at minimum channel length. Also, the gate insulator thickness and device doping profiles are optimized for both low I_{off} and high drive current (I_{on}) without approaching significant gate oxide tunneling and GIDL current (gate-induced drain leakage caused by band-to-band tunneling in the drain region due to high electric field) problems [8], [9].

The second curve is the low-power scenario, where the power-supply voltage is reduced as compared to the high-performance case at the same channel length. The goal is to lower the power dissipation per device and therefore maintain a power density in the scaled technologies which is similar to that of the starting 1 μm CMOS technology by adhering as closely as practical to CE scaling. At the same time, the speed relative to the high-performance case should not degrade more than an arbitrary 1.5×. The device design modifications and the threshold voltage choice in this case should also provide acceptable leakage current. This will be discussed in more detail in the next section.

The third curve which is given for comparison is the constant electric field (CE) scaling as proposed in [1]. The constant electric field of 5 V/μm is chosen as the starting point, since the CMOS circuits operating at 5 V using 1 μm channel lengths are fairly well optimized with respect to performance versus power dissipation. The electric field strength across the channels are sufficient to bring the operation into the saturation-velocity limited regime (for example, the average field of 5 V/μm across a single p-channel device with a mobility of 200 cm^2/Vs projects to a velocity of 10^7 cm/s). Also, any increase in vertical electric field is not very productive because parasitic series source and drain resistances would become very significant compared to the inversion layer resistances.

In the high-performance scenario of Fig. 4, the average electric field is twice as high for 0.25 μm CMOS with a 2.5 V power supply compared with 1 μm CMOS at 5 V. However, this higher electric field can be tolerated without undue reliability impact as shown in the previous section. It appears that reducing the applied voltage below the barrier height between the silicon channel and the silicon-dioxide gate insulator is helpful, even though some carriers still attain high enough energy to pass through that barrier [10]. The average electric field increases further as the channel lengths are scaled down to 0.15 μm and 0.1 μm with the power-supply voltages of 1.8 V and 1.5 V, respectively. For the 0.15 μm CMOS technology, it has been demonstrated that the CHC degradation of the nFET is maintained within an acceptable limit for 10 years operation for the minimum channel length of 0.1 μm at 1.8 V [11]. At the same time a performance improvement of 1.5× over the 0.25 μm CMOS technology has been measured. The studies of CHC effects in the 0.1 μm regime indicate that, despite the low power-supply voltage, the degradation caused by CHC cannot be ignored [12], [13]. For the high-performance 0.1 μm CMOS a power-supply voltage of 1.5 V is chosen, which achieves 2× performance improvement over the 0.25 μm CMOS [14]. This performance is limited by the threshold voltage, V_t, which is not fully scaled in order to maintain acceptable I_{off}. The gate oxide thickness for the 0.15 μm and 0.1 μm CMOS technologies are 5 nm and 3.5 nm, respectively. It has been shown that the tunneling current density in the 3.5 nm gate oxide for 1.5 V operation is 10^{13} A/μm^2, which is insignificant for most ULSI circuits and also that the gate-induced drain leakage (GIDL) current is negligible relative to the off-current at minimum channel length, and therefore does not pose any significant limitation on the device design [15].

C. Performance/Power Tradeoff and Nonscalability of the Threshold Voltage

The CMOS circuit power dissipation can be expressed as $P = KCV^2f + I_{\text{off}}V$, where K is the switching factor, C is the total load capacitance, and f is the clock frequency. The first term is the active and the second is the standby power dissipation. As CMOS is scaled to small dimensions and the power-supply voltage is reduced to maintain reliability and reasonable active power dissipation, the threshold voltage (V_t) needs to be scaled down at the same rate as the power-supply voltage in order to achieve the desired circuit switching speed. However, lowering the threshold voltage will cause substantial increases in I_{off} and the standby power [16].

The limitation which drives the nonscalability of the V_t is the turnoff behavior of the MOSFET, characterized by the inverse subthreshold slope (S) which is invariant with

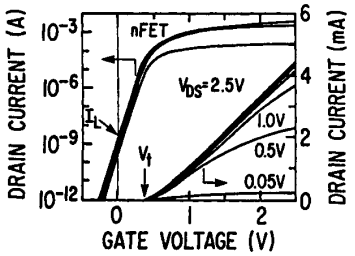


Fig. 5. 0.25 μm nFET device characteristics (W = 10 μm). Subthreshold slope = 78 mV/dec.

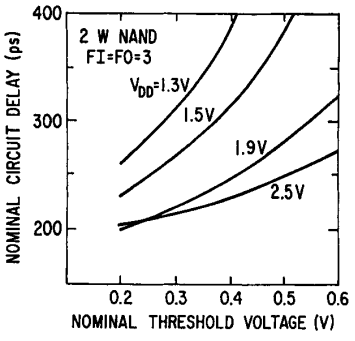


Fig. 6. 0.25 μm CMOS delay versus threshold voltage and V_{DD}.

scaling [1]. In typical scaled devices S is about 80 mV/*dec* at room temperature, as shown in Fig. 5 for a 0.25 μm nFET. At 85°C the S is about 100 mV/*dec*. Therefore, in general, for every 100 mV reduction in V_t the standby current will be increased by one order of magnitude. This exponential growth of the standby current tends to limit the threshold voltage reduction to about 0.3 V for room temperature operation of conventional CMOS circuits. It should be noted that in this paper the threshold voltage generally refers to the nominal value, extrapolated from the linear drain current versus gate voltage curve at low drain bias as shown in Fig. 5.

The effect of the threshold voltage on performance for various power-supply voltages in 0.25 μm CMOS is shown in the simulation results of Fig. 6 [17], [18]. It can be seen that as the power-supply voltage is reduced, the performance degrades significantly at higher threshold voltages and also becomes more sensitive to tolerances in V_t. Therefore, from the performance point of view, we need to reduce the threshold voltage and also improve the threshold voltage control (reduce V_t variation due to various process tolerances) as the power-supply is scaled down. It should be noted that if we scale the technology below 0.25 μm, the curves in Fig. 6 still apply if the voltages are scaled down with dimensions and the delay axis is multiplied by a constant improvement factor. The optimum V_t can be chosen for different applications, based on the tradeoff requirements between the speed and standby power.

Another critical factor in this tradeoff is the short channel effect (SCE) of the devices. SCE is the lowering of the threshold voltage at short channel lengths of a given technology relative to the threshold voltage at nominal channel length. Since the standby power is dominated by the shortest channel devices, a high SCE will force a higher nominal device V_t which will result in lower performance (Fig. 6). To some degree SCE reduces naturally in scaled devices due to reduced gate insulator thickness and increased doping levels. Other factors such as nonscalability of the band-bending and built-in junction potentials make it worse [3]. Reduction of the SCE constitutes a major part of the technology optimization and development for scaled CMOS generations. Some of the techniques to reduce SCE are summarized in the next section.

To deal with the problem of nonscalability of the threshold voltage, one possibility is to implement multiple threshold voltages on a chip, at a slightly higher processing cost. Today, dual thresholds are commonly used in DRAM chips where it is possible to raise the V_t of the array devices with a fixed body bias. For logic, low V_t devices can be used where needed for circuit functionality or higher speed, with high V_t devices for the rest of the chip. This can allow the threshold voltage of the low V_t devices to be scaled below 0.3 V. It has also been shown that by active adjustment of the substrate or well bias, including a feedback loop, one can reduce the V_t variation significantly [19]. Although these techniques do not address the fundamental problem of nonscalability of the inverse subthreshold slope, nevertheless they offer possibilities to reduce the threshold voltage without suffering from overwhelming leakage current and standby power. Unfortunately most of the system and architectural power management techniques do not address the issue of the standby power (they address the active power), unless the power management system completely shuts down all power to the chip. This is not usually practical due to the relatively long response time associated with the power management system compared to the system cycle time.

The inverse subthreshold slope of the FET's can be significantly improved (about a factor of 4), by operating the devices at 77 K, liquid nitrogen temperature [20], [21]. In this case the threshold voltages can be scaled with the power-supply voltage without leakage problems, e.g., the V_t can be 0.2 V for a 1.0 V power-supply voltage. The lower V_t along with improved carrier mobility and reduced interconnect resistance, results in significant improvement in performance and power-delay product. However, the application of 77 K CMOS will probably be limited to the highest performance systems. CMOS on silicon on insulator (SOI) offers somewhat improved subthreshold behavior as well as better performance and power-delay product [22]. This will be discussed further in the following sections.

As shown by the generalized scaling theory, the active power density will be considerably higher for the scaled devices in the high-performance scenario (Fig. 4) because of the increased electric field. The relative active power densities for several scaled CMOS technologies are shown in Fig. 7. The relative density for each technology generation is shown in parentheses. This relative density is the

Table 2 CMOS Scaling Guidelines for the Next 10 Years

	Late 1980's	1992	1995	1998	2001	2004
Supply Voltage (V)						
High Performance	5	5/3.3	3.3/2.5	2.5/1.8	1.5	1.2
Low Power	—	3.3/2.5	2.5/1.5	1.5/1.2	1.0	1.0
Lithography Resolution (μm)						
General	1.25	0.8	0.5	0.35	0.25	0.18
Gate Level for Short L	—	0.6	0.35	0.25	0.18	0.13
Channel Length (μm)	0.9	0.6/0.45	0.35/0.25	0.2/0.15	0.1	0.07
Gate Insulator Thickness (nm)	23	15/12	9/7	6/5	3.5	2.5
Relative Density	1.0	2.5	6.3	12.8	25	48
Relative Speed						
High Performance	1.0	1.4/2.0	2.7/3.4	4.2/5.1	7.2	9.6
Low Power	—	1.0/1.6	2.0/2.4	3.2/3.5	4.5	7.2
Relative Power/Function						
High Performance	1.0	0.9/0.55	0.47/0.34	0.29/0.18	0.12	0.077
Low Power	—	0.27/0.25	0.20/0.09	0.08/0.056	0.036	0.041
Relative Power/Unit Area						
High Performance	1.0	2.25/1.38	3.0/2.1	3.7/2.34	3.12	3.70
Low Power	—	0.7/0.63	1.25/0.6	1.02/0.72	0.90	1.97

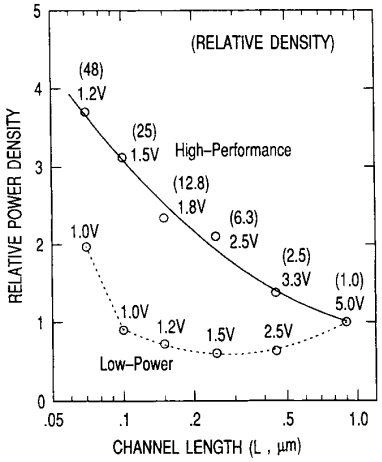


Fig. 7. Relative active power density in CMOS scaled as in Table 2. Relative density is in parentheses.

reciprocal of the square of the relative lithography ground rules for each generation (see Table 2). The upper curve corresponds to the performance driven scenario. In this case, we can see that even with reduced power-supply voltage, the power density increases significantly due to a large increase in the number of devices per unit area. However, one can choose the low power scenario, where the supply voltage is reduced more aggressively, at the expense of the performance (lower curve in Fig. 7). This scenario will be quite attractive for applications where low power and improved power-delay product is the highest priority. As an example, if we compare the 1.5 V operation versus the 2.5 V in 0.25 μm CMOS, there will be over 3.5$\times$ reduction in power (Fig. 7) with only about 30% performance degradation (Fig. 6, $V_t \simeq 0.3$ V).

In the low-power scenario, the power density of the scaled technologies goes down relative to the 1 μm CMOS technology until we reach 1.5 V (Fig. 7). This is due to the fact that in that regime the electric field does not go up substantially and is quite close to the constant electric field scaling as shown in Fig. 4, while the interconnection dimensions (which determine the density) are being scaled down less than the channel lengths. However, beyond the $L = 0.25\,\mu$m at 1.5 V point, the relative power density rises (Fig. 7) as we depart from the CE scaling (Fig. 4) because the lower limit on V_t imposes a lower limit of about 1 V for the power-supply voltage without significantly impacting speed.

A key concern in the low-power scenario is the availability of the complete chip set to make up the systems at reduced supply voltage. However, most of the problems can be overcome by various techniques to mix and match different supply voltages on the board or on the chip. Also, the need for a total low power system solution at low cost should drive the semiconductor industry to even a faster pace for offering various memory and ASIC products at reduced supply voltages.

Another key concern is the susceptibility to soft errors due to alpha particles, which is expected to increase due to reduced voltage and capacitance. This could require improved structures such as SOI to reduce the volume of junction area exposed to alpha particle hits.

D. CMOS Scaling Guideline for the Next Ten Years

The above scaling optimizations and limitations are summarized in Table 2 as a guideline for CMOS scaling over the next ten years for logic applications such as microprocessors. The key design parameters are the power-supply voltage, the lithography resolution, the channel length, and the gate insulator thickness. The more aggressively scaled device and power-supply voltages represent the high-performance and low-power scenarios discussed above. In the columns through the 90's less aggressively scaled device design points are included, representing lower cost options with correspondingly lower speed and higher power dissipation. It is assumed that the wiring pitch scales with the general lithography resolution and the relative density is the reciprocal of the square of the relative

lithography resolution. The gate level lithography requires higher resolution than the general lithography in these applications, which is in-line with the "selective scaling" that was discussed earlier. However, the critical dimension (CD) requirements of the gate level will be satisfied with the lithography tools that are developed for the density driven products and applications, i.e., DRAM, at any given generation. It is assumed that DRAM products will continue the 4× bits/chip improvement every three years. If that trend is slowed down due to productivity (cell size) and/or economics reasons, it is likely that the pace of the logic scaling proposed in this guideline will also be slowed down.

The effective channel length and the gate oxide thickness in the high-performance scenario are scaled in concert with the power-supply voltage in order to achieve optimum speed, while maintaining acceptable leakage current and reliability margins. In the low-power case, the power-supply voltage is chosen to be lower in order to reduce the power at the expense of some loss of performance. The device dimensions in the low-power scenario are assumed to be similar to the high-performance case. The reason for not using relaxed channel lengths for the low-power case is to minimize the performance degradation. This allows the speed of the low-power case in one generation to be about the same as the speed of the high-performance scenario of the previous generation, with greatly reduced power consumption.

The leakage current is mainly dominated by the devices with the shortest effective channel within the channel length distribution. Therefore, a critical parameter is the channel length tolerance which is assumed to be about 30% of the nominal channel length, leading to a gate lithography tolerance requirement of about 20% of the gate lithography resolution. The gate lithography tolerance is a very key parameter in high-performance CMOS applications since it directly affects the performance at a given leakage current requirement. This parameter is usually the most difficult and expensive feature to control in semiconductor manufacturing. The relative performance of the technology generations are arrived at through various experimental data points [4], [11], [14], [15] and projections based on the scaling principles. Many of the assumptions used in the scaling theory become less accurate at very small dimensions [3]. However, there are compensating effects which tend to make the scaling projections still reasonably accurate. One positive effect is the onset of velocity overshoot behavior which tends to increase the current in very small nFETs [23].

In general it is safe to consider that the devices will meet the long-term CHC and gate oxide reliability requirements for 10 years use even down to sub-0.1 μm channel lengths, because of the much lower operating voltage. Also, the reduction of the device series resistance by sharpening the S/D profiles will not keep pace with the power-supply voltage reduction. This results in lower electric field at the drain region and lower CHC. However, meeting the increasingly demanding yield and defect density requirements of the ultrathin gate oxides and ultrashort devices are some of the key future manufacturing challenges. It appears that fundamental limitations such as statistical dopant fluctuations [24], tunneling through the gate oxide [15], GIDL current [8], and interconnect RC delays [25] do not pose serious barriers for the scaled devices down to 0.05 μm channel length (minimum channel length, year 2004) with 2.5 nm gate oxide, operating at 1 V power-supply voltage.

Table 3 Power/Performance/Density Improvement for a RISC Processor with Scaling

	66 MHz	100 MHz	150 MHz
0.65 μm Lith. (120 mm^2)			
L = 0.45 μm, 3.6V	7.5 W		
0.5 μm Lith. (74 mm^2)			
L = 0.25 μm, 2.5 V		3.4 W	
L = 0.25 μm, 1.5 V	0.8 W		
0.35 μm Lith. (36 mm^2)			
L = 0.15 μm, 1.8 V			1.8 W
L =0.15 μm, 1.2 V		0.5 W	

For the scaled CMOS in the year 2004, a performance gain of about 7×, a density improvement of 20×, and a power/function reduction of 12× are projected (Table 2, high-performance case) compared with the present 0.6 μm technology at 5 V. However, even in the low-power scenario in the year 2004, the active power density will be higher than today's 5 V technology. The reason is the lower limit of 1 V that we imposed on the power-supply voltage due to nonscalability of the threshold voltage. It has been shown that by operating the 0.1 μm CMOS devices at power-supply voltages lower than 1 V the active power dissipation can be much reduced, but at the cost of increasing the standby power and significantly reducing the performance [26]. In the year 2004, the high-performance and the low-power scenarios exhibit about 80× and 110× improvement in the power-delay product compared with today's 5 V technology, respectively.

An example of the effectiveness of the CMOS scaling in simultaneously improving performance, power and density is shown in Table 3. In this table the operating frequency, chip size, power-supply voltage, and power dissipation of an existing high performance RISC processor and projections into more advanced scaled CMOS are presented. The 66 MHz chip with 120 mm^2 die size at 0.65 μm lithography, consumes about 7 W. The 100 MHz chip, consuming only 3.4 W, is fabricated in selectively scaled 0.25 μm CMOS technology [27] with 74 mm^2 die size, operating at 2.5 V with 3.3 V or 5 V I/O. It is estimated that by operating this chip at 1.5 V the power dissipation will be reduced to 0.8 W at 66 MHz. These numbers are in good agreement with the scaling guidelines in Table 2. It is further estimated that if this processor is fabricated in 0.35 μm lithography with 0.15 μm effective channel length, operating at 1.8 V, the performance will be improved by 50%, the die area reduced by more than 2×, and the power dissipation reduced by about 1.8×, simultaneously. Dropping the internal supply voltage to 1.2 V will further reduce the power dissipation by another 3×, while the performance will degrade by 50%.

The small size of the RISC processor if fabricated in the 0.35 μm lithography generation projected for 1998

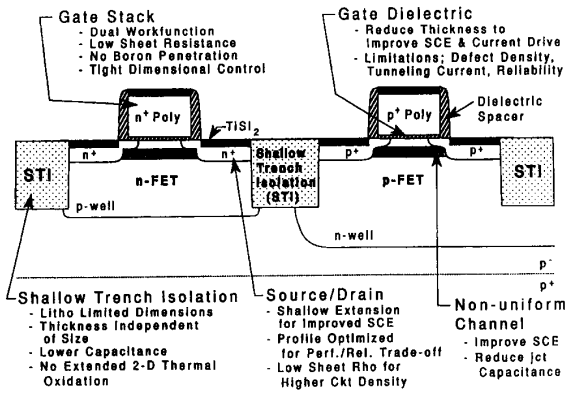


Fig. 8. Key technology elements for deep-submicron CMOS.

raises the issue of how the density improvement, projected to increase an additional 4× by 2004, will actually be utilized. Up to now logic chips have grown in size with each successive generation. If this trend is to continue, the boundaries of integration will have to be widened to take in more (or even all) of a system function onto a chip. Some of that function may require other devices and structures to provide analog, dense memory, or nonvolatile memory for example. This is a very important issue, but not one which can be further addressed here. Rather we hope the guideline for CMOS scaling given here will help determine the proper course for the next integration phase of the microelectronic era.

E. Key Technology Elements and Extensions

In order to make the above CMOS scaling a reality, several challenging technology issues must be resolved. It is assumed that the lithography and etching capability for very small dimensions will continue to be driven by the requirements for DRAM as discussed earlier. Therefore, we briefly review here the critical requirements for device and interconnection process developments.

1) Device Technology: The key technology elements for the deep submicron CMOS, excluding the interconnect, are depicted in Fig. 8. They are divided into four modules: Gate stack and gate dielectric, source/drain, isolation, and channel profile. The main requirements for each group are summarized in Fig. 8 [31], [28] and will not be repeated here. Some of the key features are performance driven, e.g., more abrupt source drain (S/D) profiles to lower device series resistance, thinner gate dielectric to increase drive current and improve SCE, and silicided gate to reduce the RC delay for wide devices. It is noteworthy that the reduction of the gate dielectric thickness mainly depends on defect density requirements rather than tunneling current. It has been shown that the defect density is a function of the electric field and that it goes up at higher operating fields [9], [29]. It can be projected that in order to maintain defect densities below 1/cm^2, the oxide field should be below 5 MV/cm in the oxide thickness range below 5 nm [9]. This condition is satisfied in the scaled devices in Table 2.

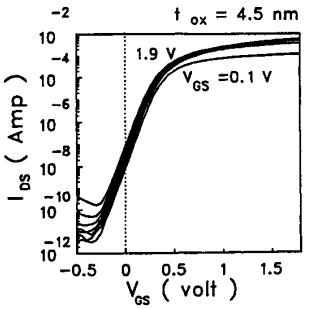


Fig. 9. Turn-on characteristics, drain current versus gate voltage with drain voltage as a parameter (0.3 V steps), for an $L = 0.1$ μm nFET in a 0.15 μm CMOS process ($W = 10$ μm).

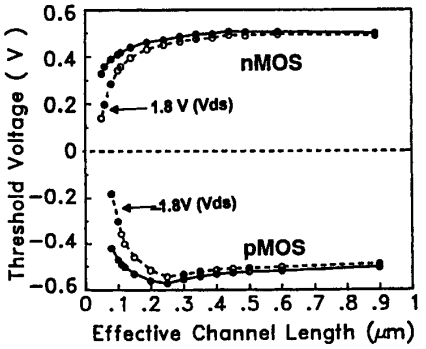


Fig. 10. Threshold voltage roll-off in high-performance 0.15 μm CMOS versus channel length.

Some other technology modules are mainly density driven, e.g., shallow trench isolation (STI), and silicided S/D junctions. STI results in denser circuits, offering lithography limited isolation widths by eliminating bird's beak and offering improved planarity for the critical gate definition step [30]. It should be mentioned that STI also improves circuit performance due to minimizing the perimeter junction capacitance by offsetting the field doping away from the junction edge. The silicided S/D junctions improve the density by reducing the number of contacts required to the S/D area, and therefore freeing up wiring tracks. The main silicidation issues for the reduced dimensions of scaled CMOS are discussed elsewhere [31], [32].

Reducing SCE is of particular importance in scaled CMOS technologies with reduced power-supply voltage, as discussed previously. It has been demonstrated that by using nonuniform channel doping profiles and source/drain extensions with reverse doping preamorphization [33], excellent SCE can be achieved for both nFET and pFET devices as shown in Fig. 9 and Fig. 10 [34]. Less than 200 mV threshold voltage roll-off is achieved down to 0.1 μm at high drain voltage (Fig. 10). Using these devices, high performance CMOS circuits, exhibiting delay per stage of 35 ps at 1.8 V at $L = 0.15$ μm for unloaded, and 200 ps loaded (FI = FO = 3, $C_{\text{wire}} = 240$ fF) have been demonstrated as shown in Fig. 11 [34].

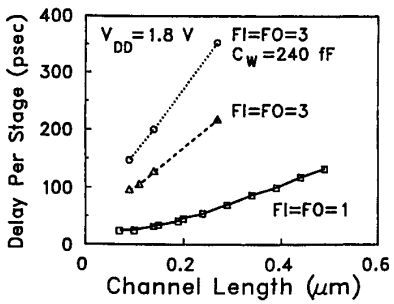


Fig. 11. Measured delay versus channel length for the high-performance $L = 0.15$ μm CMOS technology ($W_n = W_p/2 = 15$ μm).

2) Interconnect Issues for High-Performance Applications: With the scaled devices discussed above, CMOS microprocessor speeds should reach in excess of 500 MHz (2 ns cycle time). As the high-end microprocessors get implemented in CMOS technology [35], one needs to pay special attention to the interconnect RC delays, so that the chip performance is not gated by the wiring delay [1], [25]. It has been discussed that due to the nonscalability of the interconnect delays caused by the finite wire resistance, a hierarchy of interconnect levels should be used. First, there are the "short" wires (less than a few hundred microns long) which serve the vast majority of the chip interconnects, and are responsible for the chip wirability by providing a sufficient number of wiring channels. These wires should scale with the lithography as discussed earlier in the scaling guideline. Second, there is a need for "long" wires, where density is secondary to delay considerations [25]. These interconnections were a part of the package in earlier days, but with increased level of integration they are now on the chip as we get closer to the "high-performance systems on a chip" paradigm. These wires run between distant parts of the chip, connecting various functional blocks. The signal propagation delay on the "long" wires should be maintained at a small fraction of the processor cycle time. Therefore, these wires cannot shrink with the rest of the chip dimensions.

This wiring hierarchy is depicted in Fig. 12 [25]. In the top two levels which will be used as long wires, the capacitance per unit length stays constant, while resistance decreases proportionally with the increase in wire cross section. These wires are referred to as "fat" wires by Sai-Halasz [25]. One consequence of having low RC wires is that we will observe transmission line behavior not only on the package, but on the chips as well. For illustration, on a 15 mm long wire the signal flight time cannot be less than 105 ps, which is longer than the switching time of drivers in the scaled technologies below 0.25 μm proposed in the guideline. When the input of a wire is driven with a signal faster than the travel time down that line, we reach the regime where the nonideal transmission line characteristics of the present-day interconnect schemes must be taken into account.

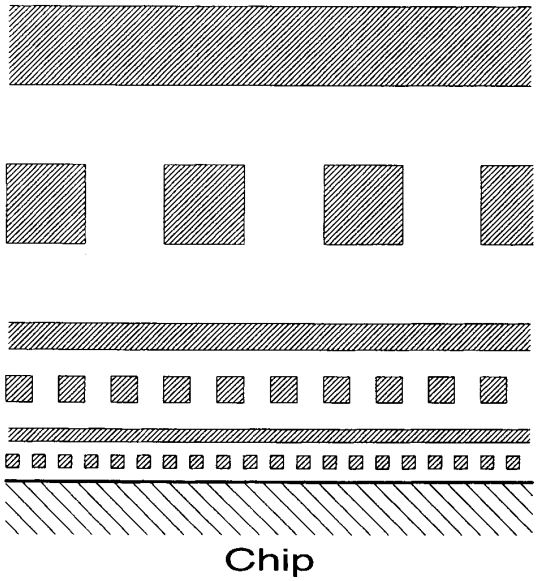


Fig. 12. Example of wiring scheme needed by future high-performance processors to minimize delays due to wire resistance. Three x-y wiring planes are shown in cross section.

Due to the stringent RC requirements for the "fat" wires, it is unlikely that the large width and height of these wires can be significantly reduced by using new conductor and insulator materials with lower resistivity and dielectric constant, respectively. This does not diminish the importance of exploring new interconnect materials with superior properties relative to today's Al/SiO_2 system. It should be mentioned that in order to minimize the negative impact of the "fat" wires on the chip wiring density, they should not replace the "short" wiring levels; rather, they should be added to the already existing interconnect scheme. Due to this and also the complicated processing associated with the etch and fill of high aspect ratio via holes, the "fat" wires can potentially increase the processing cost significantly (performance/cost tradeoff). The key elements of the interconnect technology are not discussed here, and can be found elsewhere [36], [37].

Another important function of the interconnect technology is distribution of power to the processing elements. As the chip's complexity grows and the speed increases, supplying peak currents from outside the chip (or even across the chip) becomes very difficult because of inductance effects. Therefore, it may be necessary to provide large amounts of decoupling capacitance distributed throughout the chip to minimize power-supply noise. This is another example of integrating a function which was previously a part of the package.

3) Silicon On Insulator (SOI): Additional significant improvements in power and/or performance can be achieved by implementing scaled CMOS on silicon on insulator (SOI). A schematic cross section of CMOS on SOI is shown in Fig. 13 [38]. The performance improvement of SOI compared to bulk CMOS is mainly due to the reduction

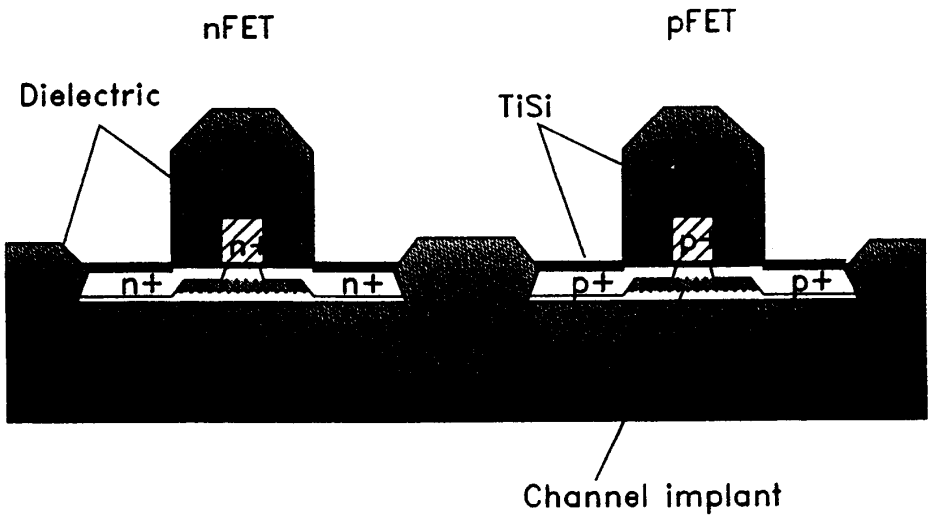


Fig. 13. 0.1 μm CMOS on SOI schematic cross section.

of parasitic capacitances and body effect. Also, in partially depleted device designs, the floating body effect can give rise to a sharper subthreshold slope ($S <$ 60 mV/dec) at high drain bias, which effectively reduces the threshold voltage and can actually improve the performance at a given standby leakage current [39]. A major issue here is reduction of the breakdown voltage due to the bipolar action, particularly in the nFET. To reduce the bipolar gain in order to maintain acceptably high breakdown voltage both for normal operation and for burn-in, S/D extensions with halo can be employed [38]. CMOS on SOI can extend the bulk CMOS performance limits as well as improving the performance at a given fab generation. Performance improvements in the range of 1.5×–2.5× have been reported for various CMOS on SOI circuits, relative to CMOS on bulk devices with the same lithography [22], [40]. In addition, CMOS on SOI offers significant reduction in the soft error rate, latch-up elimination, and simpler isolation which results in reduced wafer fabrication steps.

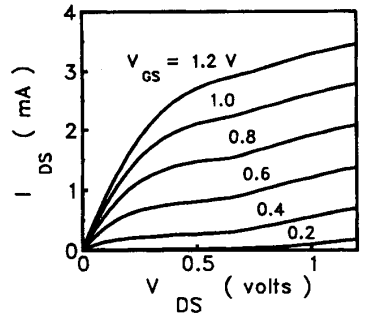
The main challenges and shortcomings of SOI are the availability of low-cost wafers with low defect density, floating body effects on the device and circuit operation, and heat dissipation through the buried oxide. Progress in starting SOI wafers has been made rapidly in the last few years [41]. The process which uses implanted oxygen to form an SiO_2 layer beneath a thin silicon layer (SIMOX) is emerging as a viable process after years of development. The other major approach, which uses bonding of two oxidized wafers and etchback of all but a thin layer of one wafer (BESOI), has also shown capability to make thin uniform layers in the required range of 0.1–0.2 μm thick. Defect density has also been greatly reduced over the last few years and is approaching the same level as that of bulk substrates.

The impact of the SOI shortcomings becomes significantly less as the supply voltage is reduced below 2.5 V, while the impact of the SOI benefits becomes stronger [42]. Therefore, SOI is an excellent candidate for low power/low voltage applications. The I-V characteristics of the nondepleted 0.1 μm devices shown in Fig. 14 are reasonably free of kink effect in the range where they will be operated [39]. By using CMOS on SOI, more than 3× reduction in power-per-stage can be achieved at the same performance and channel length, compared with CMOS on bulk as shown in Fig. 15 for L = 0.15 μm unloaded ring oscillators [42]. Also an SRAM access time of 3.5 ns has been demonstrated at 1.0 V in an experimental 512 Kb circuit [39].

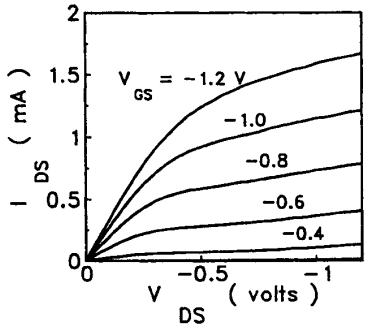
III. Summary and Conclusions

Scaled CMOS technology is ideally suited as the engine for both tomorrow's high performance systems and for the coming low power revolution. Within the next decade, there will be an explosive growth of capability in silicon chips, made possible by CMOS scaling. This growth will affect all aspects of human life as the integration of high performance systems on a single chip (as powerful as today's supercomputers) becomes a reality. In this article, a guideline for CMOS scaling over the next 10 years has been presented, with emphasis on the optimization of high-performance and low-power scenarios aimed at logic applications such as microprocessors. After considering key device and technology bottlenecks and various nonscalable elements, it is projected that the performance, density, and active power dissipation will all improve dramatically as scaling proceeds along the path presented in the guideline.

In the year 2004, using sub-0.1 μm devices, speed improvement of about 7×, density improvement of about 20×, and power/function reduction of more than 10× are expected relative to today's 5 V technology at 0.6 μm. The power-delay product will improve by 80× and 110× for high-performance and low-power scenarios, respectively.



(a)



(b)

Fig. 14. Device characteristics of 0.1 μm CMOS on SOI (W =10 μm). (a) nFET, (b) pFET.

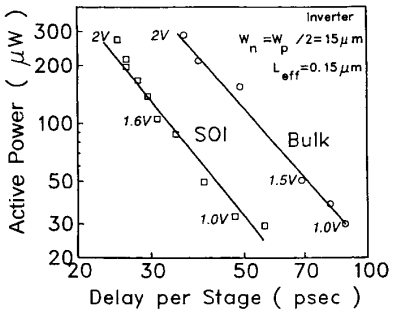


Fig. 15. Measured power versus delay for L = 0.15 μm SOI and bulk CMOS with varying voltage.

Reduction of the power-supply voltage is a key element in future CMOS scaling. The increasing demands for reduced power dissipation, as well as the eroding reliability margins, will result in much more frequent power-supply voltage changes in the industry than in the past "constant voltage scaling" era (5 V). Even with the reduction of the voltage, the active power density will grow due to the high rate of density and performance improvements. Standby power will also grow significantly as the threshold voltage is reduced. New design techniques and power management aimed at reduction of active power and possibly standby power are needed [43]. The development of standards and practices which allow the coexistence of chips with multiple supply voltages at the system level, without compromising speed or cost (e.g., low voltage swing I/O standards) are of high priority. However, it is likely that the great increase in density will cause more and more of the system to be swept up and integrated into a single chip, certainly for the most pervasive, low-cost, and portable applications. Improvements in the interconnection technology for long cross chip wires are necessary to keep up with the advances in device speed.

Most of the projected progress in device technology has already been demonstrated in the laboratory. The work that has been done on SOI is also very promising although it has not yet been readied for products. There are, of course, many concerns about manufacturability for the path shown in the roadmap, including the vital issue of lithography tools. In summary, many practical challenges remain to be faced in the upcoming decade, but the benefits are sure to be very rewarding.

ACKNOWLEDGMENT

The authors wish to acknowledge M. Hakey, T. H. Ning, and M. R. Polcari for their contributions to this work and many very helpful 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 MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circ.*, vol. SC-9, pp. 256–268, May 1974.

[2] B. Hoeneisen and C. A. Mead, "Fundamental limitations in microelctronics -I. MOS technology," *Solid State Electron.*, vol. 15, pp. 819–829, 1972.

[3] G. Baccarani, M. R. Wordeman, and R. H. Dennard, "Generalized scaling theory and its application to a 1/4 micron MOSFET design," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 452–462, Apr. 1984.

[4] B. Davari *et al.*, "A high performance 0.25 μm CMOS technology," *IEDM Tech. Dig.*, pp. 56–59, 1988.

[5] S. Ogura, P. J. Tsang, W. W. Walker, D. L. Critchlow, and J. F. Shepard, "Design and characterization of the lightly doped drain (LDD) insulated gate field effect transistor," *IEEE Trans. Electron Devices*, vol. ED-27, pp. 1359–1367, Aug. 1980.

[6] B. Davari *et al.*, "A high performance 0.25 μm CMOS technology: Part II-Technology," *IEEE Trans. Electron Devices*, vol. 39, pp. 967–975, Apr. 1992.

[7] K. K. Ng and W. T. Lynch, "The impact of intrinsic series resistance on MOSFET scaling," *IEEE Trans. Electron Devices*, vol. ED-34, pp. 503–511, Mar. 1987.

[8] T. Y. Chan, J. Chen, P. K. Ko, and C. Hu, "The impact of gate-induced drain leakage current on MOSFET scaling," *IEDM Tech. Dig.*, pp. 718–721, 1987.

[9] R. Moazzami and C. Hu, "Projecting gate oxide reliability and optimizing reliability screens," *IEEE Trans. Electron Devices*, vol. 37, pp. 1643–1650, July 1990.

[10] J. E. Chung, M. C. Jeng, J. E. Moon, P. K. Ko, and C. Hu, "Low-voltage hot-electron currents and degradation in deep-submicron MOSFET's," *IEEE Trans. Electron Devices*, vol. 37, pp. 1651–1657, July 1990.

[11] G. G. Shahidi *et al.*, "A high performance 0.15 μm CMOS," *1993 Symp. on VLSI Technology*, Kyoto, Japan, pp. 93–94.

[12] M. Dutoit *et al.*, "Experimental study of electron heating in 0.1 μm nMOSFETs," *1993 Symp. on VLSI Tech.*, Kyoto, Japan, pp. 35–36.

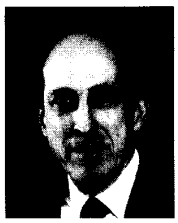
[13] T. Mizuno *et al.*, "Hot-carrier effects in 0.1 μm gate length CMOS devices," *IEDM Tech. Dig.*, pp. 695–698, 1992.

[14] Y. Taur *et al.*, "High performance 0.1 μm CMOS devices with 1.5 V power supply," *IEDM Tech. Dig.*, pp. 127–130, 1993.

[15] Y. Mii *et al.*, "High performance 0.1 μm nMOSFET's with 10 ps/stage delay (85 K) at 1.5 V power supply," *1993 Symp. on VLSI Tech.*, Kyoto, Japan, pp. 91–92.
[16] E. J. Nowak, "Ultimate CMOS ULSI performance," *IEDM Tech. Dig.*, pp. 115–118, 1993.
[17] R. H. Dennard, "Power-supply considerations for future scaled CMOS systems," *1989 Symp. on VLSI Tech., Syst. and Applications*, Taipei, Taiwan, pp. 188–192.
[18] B. Davari, R. H. Dennard, and G. G. Shahidi, "CMOS technology for low voltage/low power applications," *Proc. IEEE 1994 CICC*, pp. 3–10.
[19] T. Kobayashi and T. Sakurai, "Self-adjusting threshold-voltage scheme (SATS) for low-voltage high-speed operation," *Proc. IEEE 1994 CICC*, pp. 271–274.
[20] F. H. Gaensslen, V. L. Rideout, E. J. Walker, and J. J. Walker, "Very small MOSFET's for low-temperature operation," *IEEE Trans. Electron Devices*, vol. ED-24, pp. 218–229, Mar. 1977.
[21] J. Y.-C. Sun, Y. Taur, R. H. Dennard, and S. P. Klepner, "Submicrometer-channel CMOS for low-temperature operation," *IEEE Trans. Electron Devices*, vol. ED-34, pp. 19–27, Jan. 1987.
[22] G. G. Shahidi *et al.*, "Fabrication of CMOS on ultrathin SOI otained by epitaxial lateral overgrowth and chemical-mechanical polishing," *IEDM Tech. Dig.*, pp. 587–590, 1990.
[23] G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, S. Rishton, and E. Ganin, "High transconductance and velocity overshoot in nMOS devices at the 0.1 μm gate-length level," *IEEE Trans. Electron Device Lett.*, vol. 9, pp. 464–466, Sept. 1988.
[24] H. S. Wong and Y. Taur, "Three dimensional atomistic simulation of discrete random dopant distribution effects in sub-0.1 μm MOSFET's," *IEDM Tech. Dig.*, pp. 705–708, 1993.
[25] G. A. Sai-Halasz, "Performance trends in high-end processors," *IEEE Proc.*, vol. 83, pp. 20–36, Jan. 1995.
[26] Y. Mii *et al.*, "An ultra-low power 0.1 μm CMOS," *1994 Symp. on VLSI Tech.*, Hawaii, pp. 9–10.
[27] C. Koburger *et al.*, "Simple, fast, 2.5 V CMOS logic with 0.25 μm channel lengths and damascene interconnect," *1994 Symp. on VLSI Tech.*, Hawaii, pp. 85–86.
[28] B. Davari, "Low voltage/low power device technologies," *IEDM Short Course*, 1993.
[29] C. Hu, "Future CMOS scaling and reliability," *Proc. IEEE*, vol. 81, p. 682, May 1993.
[30] B. Davari *et al.*, "A new planarization technique, using a combination of RIE and chemical mechanical polish (CMP)," *IEDM Tech. Dig.*, pp. 61–64, 1989.
[31] ——, "Shallow junctions, silicide requirements and process technologies for sub-0.5 μm CMOS," *Proc. 22 ESSDERC*, p. 649, 1992.
[32] J. B. Lasky, J. S. Nakos, O. J. Cain, and P. J. Geiss, "Comparison of transformation to low-resistivity phase and agglomeration of $TiSi_2$ and $CoSi_2$," *IEEE Trans. Electron Devices*, vol. ED-38, pp. 262–269, Feb. 1991.
[33] B. Davari, E. Ganin, D. Harame, and G. A. Sai-Halasz, "A new pre-amorphization technique for very shallow p^+/n junctions," *1989 Symp. on VLSI Tech.*, Kyoto, Japan, pp. 27–28.
[34] G. G. Shahidi *et al.*, "Indium channel implant for improved short-channel behavior of submicrometer nMOSFET's," *IEEE Electron Device Lett.*, vol. 14, pp. 409–411, Aug. 1993.
[35] A. Masaki, "Possibilities of CMOS mainframe and its impact on technology R&D," *1991 Symp. on VLSI Technology*, Oiso, Japan, pp. 1–4.
[36] C. Kaanta *et al.*, "Submicron wiring technology with tungsten and planarization," *IEDM Tech. Dig.*, pp. 209–212, 1987.
[37] F. White *et al.*, "Damascene stud local interconnect in CMOS technology," *IEDM Tech. Dig.*, pp. 301–304, 1992.
[38] G. G. Shahidi *et al.*, "A room temperature 0.1 μm CMOS on SOI," *1993 Symp. on VLSI Tech.*, Kyoto, Japan, pp. 27–28.
[39] ——, "SOI For a 1-volt CMOS technology and application to a 512 Kb SRAM with 3.5 ns access time," *IEDM Tech. Dig.*, pp. 813–816, 1993.
[40] A. Kamgar *et al.*, "Ultra-high speed CMOS circuits in thin SIMOX films," *IEDM Tech. Dig.*, pp. 829–832, 1989.
[41] G. W. Cullen, M. T. Duffy, and A. C. Ipri, "Thirty years of silicon on insulators: do trends emerge?," in *Silicon on Insulator Tech. and Devices, Proc. ECS*, vol. 94-11, pp. 5–15, 1994.
[42] G. G. Shahidi, T. H. Ning, R. H. Dennard, and B. Davari, "SOI for low-voltage and high-speed CMOS," *Extended abstracts, 1994 Int. Conf. on Solid-State Devices and Materials*, Yokahama, Japan, pp. 265–267.
[43] A. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid State Circ.*, vol. 27, pp. 473–484, Apr. 1992.

Bijan Davari (Senior Member, IEEE) was born in Tehran, Iran, in 1954. He received his B.S. degree in electrical engineering in 1977 from Arya Mehr University of Technology (Sharif), Tehran, Iran, and the M.S. and Ph.D. degrees in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1984.

He joined IBM research division, Thomas J. Watson Research Center, Yorktown Heights, NY, in 1984. Since then he worked on various aspects of scaled CMOS and BiCMOS technologies, including device scaling and process integration. He defined and developed a selectively scaled 0.25 μm CMOS technology at 2.5 V, demonstrating significant performance and power reduction improvement over previous 0.5 μm CMOS technologies at 3.3 V. This work has set the direction and the supply voltage for the post 3.3 V CMOS generations. He is presently the senior manager of the Advanced Logic and SRAM Development in IBM's Semiconductor Research and Development Center (SRDC). His department's activities include the development of 0.1 μm CMOS for logic and SRAM products and SOI and NVRAM. He has authored and coauthored over 60 publications in various aspects of semiconductor devices and technology.



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

He then joined IBM Research Division where his early experience included the study of new devices and circuits for logic and memory applications, and the development of advanced data communication techniques. Since 1963 he has been at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, where he has been involved in microelectronics research and development from the early days onward. His primary work has been in field-effect transistors (FET's) and integrated digital circuits using these devices. In 1967 he invented the dynamic RAM memory cell used in most all computers today. With others, he developed the concept of FET scaling in 1972.

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

Ghavam G. Shahidi, photograph and biography not available at time of publication.