CMOS scaling into the 21st century: 0.1 μm and beyond

by Y. Taur
Y.-J. Mii
D. J. Frank
H.-S. Wong
D. A. Buchanan
S. J. Wind
S. A. Rishton
G. A. Sai-Halasz
E. J. Nowak

This paper describes the design, fabrication, and characterization of 0.1-μm-channel CMOS devices with dual n+/p+ polysilicon gates on 35-Å gate oxide. A 2× performance gain over 2.5-V, 0.25-μm CMOS technology is achieved at a power supply voltage of 1.5 V. In addition, a 20× reduction in active power per circuit is obtained at a supply voltage <1 V with the same delay as the 0.25-μm CMOS. These results demonstrate the feasibility of high-performance and low-power room-temperature 0.1-μm CMOS technology. Beyond 0.1 μm, a number of fundamental device and technology issues must be examined: oxide and silicon tunneling, random dopant distribution, threshold voltage nonscaling, and interconnect delays. Several alternative device structures (in particular, low-temperature CMOS and double-gate MOSFET) for exploring the outermost limit of silicon scaling are discussed.

1. Introduction

The evolution of MOSFET technology has been governed mainly by device scaling [1] over the past twenty years. One of the key questions concerning future ULSI technology is whether MOSFET devices can be scaled to 0.1-μm channel length and beyond for continuing density and performance improvement [2]. A number of device and technology issues will ultimately determine the limit of room-temperature scaling. Among the device issues are choice of power supply and threshold voltages versus active power and off-current requirements, control of short-channel effect, and hot-carrier reliability. Among the technology issues are ultrathin gate oxide, p+-polysilicon gate for surface-channel p-MOSFET, shallow source–drain junctions with low series resistance, and sub-0.2-μm lithography.

In ideal constant-field scaling, both the power supply and threshold voltages should scale linearly with channel length. However, because of subthreshold nonscaling, the threshold voltage cannot be reduced without limit. **Figure 1** shows the trend of power supply voltage, threshold voltage, and gate oxide thickness scaling versus channel length [3–5] from a mature 1-μm CMOS technology to a projected 0.1-μm CMOS technology. When the channel length is scaled down, the power supply voltage must be reduced as well to keep the device power and field (reliability) in reasonable limits. On the other hand, the threshold voltage has not been scaled in proportion to the power supply voltage. This is because the subthreshold slope, a measure of the transistor turn-off rate versus gate voltage, is largely driven by thermally activated diffusion

0018-8646/95/$3.00 © 1995 IBM

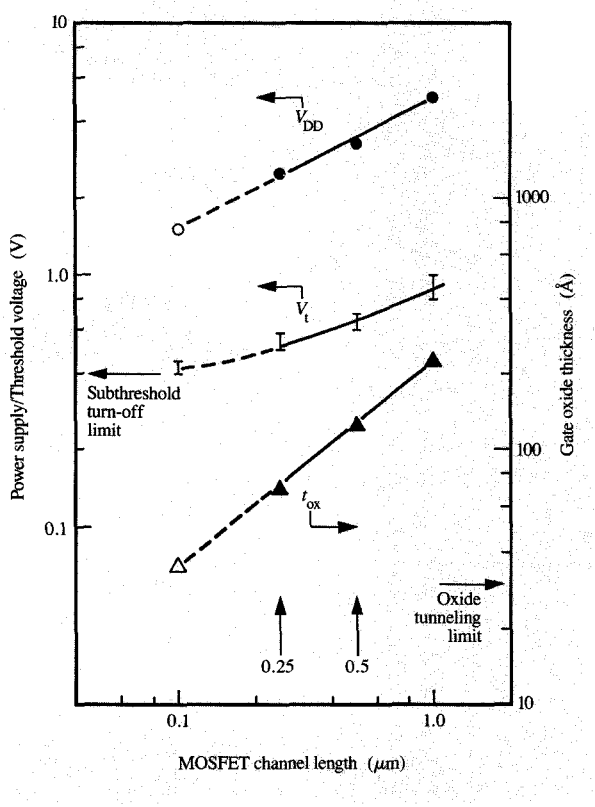

Figure 1

Power supply, threshold voltage, and gate oxide thickness trends vs. channel length for CMOS technologies from 1 μm to 0.1 μm.

and is independent of power supply voltage and channel length. In fact, for room-temperature technologies, a threshold voltage (V_t) of about 0.4 V is required, in which half (~0.2 V) is the minimum value for turning the device off, and the other half (~0.2 V) accounts for tolerances from short-channel effect and chip temperature (25°C to 85°C). Such a minimum V_t also implies a minimum power supply voltage (V_{DD}) of 1.5 V or so, since CMOS circuit delays increase rapidly when the V_t/V_{DD} ratio exceeds 1/4 [6]. Another limit on device scaling comes from gate oxide tunneling. Gate oxide thickness must be scaled down with channel length, as shown in Figure 1, to keep two-dimensional effects such as short-channel effect under control. When the gate oxide becomes thinner than 40 Å, direct quantum-mechanical tunneling occurs for voltages below the Si/SiO_2 barrier height, 3.1 eV [7]. These limits will be approached at the 0.1-μm CMOS generation.

The gate lithography resolution required for the 0.1-μm-channel CMOS discussed in this paper is in the range of 0.15–0.20 μm. Other lithography dimensions, including back-end-of-line, are assumed to be 0.25 μm (0.5-μm pitch). At present, there is no manufacturing tool capable of patterning gates smaller than 0.2 μm. This is largely a cost issue. Electron-beam lithography can easily define 0.1-μm gates, although its throughput is low. Optical lithography using an excimer laser stepper with phase-shift mask is projected to a linewidth resolution of 0.20 μm [8]. A number of research and development groups are working on X-ray lithography [9], which, in principle, can provide high-throughput 0.15-μm patterning if the cost can be contained.

In Section 2, the design, fabrication, and characterization of high-performance and low-power 0.1-μm-channel CMOS devices are described. Section 3 addresses various factors which may limit further device scaling: oxide and silicon tunneling, random dopant distribution, threshold voltage nonscaling, and interconnect delays. Section 4 discusses a number of novel material and device structures beyond 0.1-μm CMOS: SiGe, SOI, low-temperature CMOS, and double-gate MOSFET, which may bring us to the outermost limit of silicon device scaling. Section 5 concludes the paper.

2. 0.1-μm CMOS

• *Device design*

A key issue in 0.1-μm CMOS design is the choice of power supply and threshold voltages which ultimately determine the power and performance of CMOS circuits. In general, the active power of a CMOS chip is given by

$$P_{act} = (C_{sw}V_{DD}^2/2)f, \quad (1)$$

where C_{sw} is the total node capacitance being switched (either up or down) in a clock cycle, V_{DD} is the power supply voltage, and f is the clock frequency. On the other hand, the standby power of a CMOS chip is given [10] by

$$P_{off} = W_{tot}V_{DD}I_{off} = W_{tot}V_{DD}I_0\exp(-qV_{t,wc}/mkT), \quad (2)$$

where W_{tot} is the total device width having a V_{DD} drop across, I_{off} is the worst-case off-current per unit width at 85°C, I_0 is the current per unit width at threshold voltage (of the order of 10 μA/μm for 0.1-μm devices), m is a dimensionless factor typically ≈1.4, and $V_{t,wc}$ is the worst-case threshold voltage at 85°C, which is lower than the nominal room-temperature threshold voltage, V_t, by about 200 mV because of the short-channel effect and the temperature difference. To keep both active and standby power within reasonable limits, one needs to keep V_{DD} low and V_t high. However, the delay of most conventional CMOS circuits is a monotonically increasing function of V_t/V_{DD}, which increases rapidly when $V_t/V_{DD} > 1/4$ [6]. It is, therefore, important to choose optimum values of V_{DD} and V_t for a critical balance between circuit performance and chip power.

The performance–power trade-off is illustrated in **Figure 2**, where constant delay, constant active power, and constant standby power contours are plotted in a threshold voltage–power supply design plane. Both the delay and power are normalized to a reference set by 2.5-V, 0.25-μm CMOS devices (1X) [5]. For calibration, the active power of a 0.25-μm CMOS microprocessor is in the range of 5–50 W; the standby power is 10–100 mW; and the clock frequency is 100–400 MHz. The relative power values for 0.1-μm CMOS are calculated assuming no increase in the number of circuits and a factor of 4 shrinkage in the device area (from finer lithography). In general, the active power increases toward higher V_{DD} roughly as V_{DD}^2, while the standby power increases exponentially toward lower V_t as $\exp(-qV_t/mkT)$. The delay, on the other hand, increases toward higher V_t and lower V_{DD} until limited by tolerance considerations, $V_t/V_{DD} \leq 0.65$, indicated by the thick dashed line in Figure 2. For high-performance design, a power supply voltage in the range of 1.2–1.6 V is suitable for achieving a 2× performance gain over 0.25-μm CMOS with moderate reductions in both active and standby power per circuit [11]. This corresponds to a clock frequency in the range of 200–800 MHz, depending on circuit design and chip architecture, for microprocessors using 0.1-μm CMOS. For low-power design, a power supply in the range of 0.6–1.0 V can be used to achieve a 15–30× reduction in active power per circuit while still maintaining the same performance as 0.25-μm CMOS [6]. If the system can tolerate a higher standby power, more reduction in active power is possible by operating at lower V_{DD} and V_t. For high-function 0.1-μm CMOS chips in which the number of circuits increases by a factor of 4 over 0.25-μm CMOS (for the same chip size), all the power values in Figure 2 must be multiplied by 4, making the power trade-off a more important issue.

The above design trade-offs did not consider hot-electron reliability, which has been a major constraint on CMOS power supply voltage above 2.5 V. As the voltage is scaled to 1.5 V and below, however, hot-electron reliability should no longer be a limiting factor, since the average carrier energy is below the thresholds for most high-field effects, e.g., Si–SiO_2 barrier height (3.1 eV), interface state/trap generation (2.5 eV), and impact ionization (1.6 eV). For 0.1-μm CMOS technology, therefore, the power supply voltage will be limited primarily by active power considerations, as discussed above.

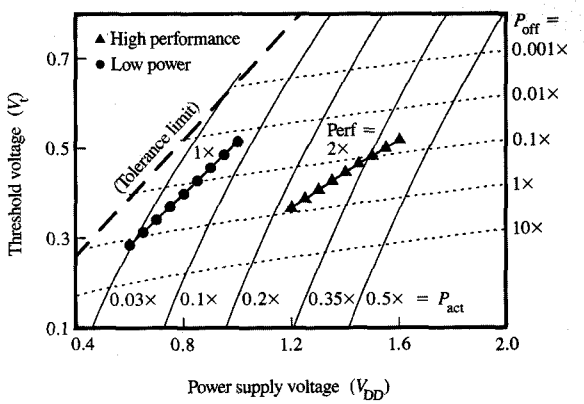

Figure 2

Constant active power/circuit (solid lines), constant standby power/circuit (thin dashed lines), and constant performance (triangles: 2×; dots: 1× 0.25-μm CMOS) contours in a threshold voltage–power supply design plane. The thick dashed line on the upper left indicates the limit imposed by V_t and V_{DD} tolerances.

Table 1 summarizes the design parameters of a high-performance 0.1-μm CMOS [11] device. The gate oxide thickness and source–drain junction depth are aggressively scaled to 35 Å and 400–600 Å, respectively. The channel doping design is that of a retrograde type [12], which, for a given threshold voltage, allows higher subsurface doping for control of short-channel effect.

Table 1 Device parameters for high-performance 0.1-μm CMOS.

	Device parameters	*n-MOSFET*	*p-MOSFET*
Design	Power supply voltage (V)	1.5	
	Gate oxide thickness (Å)	35	
	Effective channel length (μm)	0.1	
	Threshold voltages (V)	±0.4	
Experimental	Source–drain extension depth (Å)	400	600
	Subthreshold slope (mV/dec.)	90	90
	Source–drain series resistance (Ω-μm)	250	700
	Saturation transconductance (mS/mm)	550	320
	Current-gain cut-off frequency (GHz)	100	40

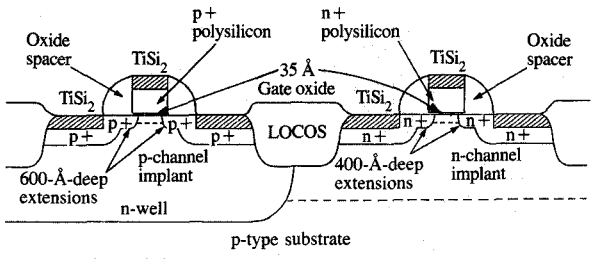

Figure 3

Schematic cross section of 0.1-μm CMOS devices.

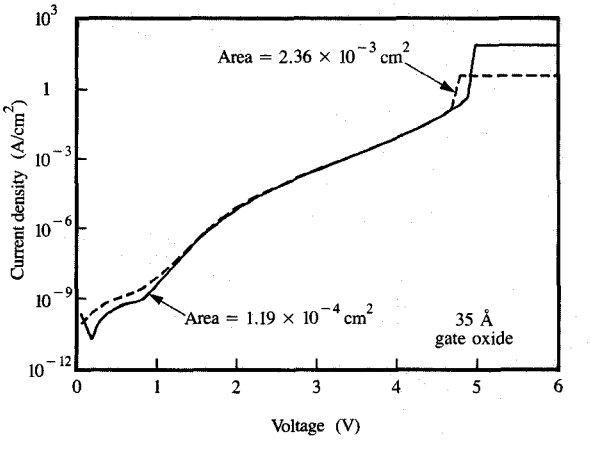

Figure 4

Measured tunneling current vs. gate voltage for two 35-Å MOS capacitors of different areas.

• *Process*

A schematic cross section of the 0.1-μm CMOS device is shown in **Figure 3**. The fabrication process includes five e-beam lithographic levels and eight optical levels (for implant block-out). The 35-Å gate oxide is grown at 750°C in dry oxygen with HCl. E-beam lithography is carried out on an ultrahigh-resolution vector scan system with a thermal field-emission source [13]. The resist for the gate level is an epoxy–novolak negative resist which has superior resolution and contrast as well as good resistance to reactive ion etching of polysilicon gates. A 0.22-μm-thick resist film is exposed at a dose of 3.5 μC/cm^2 on a 400-μm × 400-μm exposure field. The polysilicon gate etch process is optimized to achieve vertical sidewall profiles by an HBr/Cl_2 reactive ion etch chemistry. In order to stop on the 35-Å gate oxide, a high-selectivity (>100) etch is employed so that less than 10 Å of oxide is consumed.

To fabricate 0.1-μm p-MOSFETs with acceptable short-channel effects, a p+-polysilicon gate is required for surface channel operation. A common problem with p+-polysilicon is boron penetration through the thin gate oxide into the channel region, modifying the threshold voltage. In the 0.1-μm CMOS process, p+-polysilicon gates are doped by low-energy boron ion implantation and rapid thermal annealing in an argon ambient. No significant boron penetration, if any, from p+-polysilicon was observed, since the C–V flatband voltage, 0.95 V, is within 100 mV of that expected from the p+-polysilicon work function. The C_{inv}/C_{max} ratio is close to unity, indicating negligible gate-depletion effects [14]. **Figure 4** shows the tunneling current and breakdown voltage of the 35-Å oxide. The breakdown voltage is 4.7 V, corresponding to an oxide field of 11 MV/cm. The tunneling current distribution is uniform, as indicated by the nearly identical current densities from two MOS capacitors of different areas. For a 1.5-V power supply, the tunneling current is less than 10^{-14} A/μm^2, well within VLSI specifications.

When the MOSFET channel length is scaled down, both the gate oxide thickness and the source–drain junction depth must be scaled down as well to keep 2D effects such as short-channel effect under control. One of the main difficulties in 0.1-μm MOSFETs is forming an ≈500-Å-deep source–drain junction and making a low-resistance silicide contact to it. Junction leakage and/or contact resistance are common problems, since a significant layer of doped silicon is consumed in the silicide process [15]. This problem is avoided by using the source–drain extension structure in Figure 3. Shallow (≈500 Å) p+ (or n+) source–drain extensions are used in conjunction with deeper p+ (or n+) source–drain regions implanted after thick oxide spacers. This decouples the shallow extension depth from the deep junction required for the $TiSi_2$ process. A medium-dose, counter-doping implant (halo) is made with the extension implant to increase the doping level in short-channel devices and to suppress short-channel effect [16]. After source–drain implant and anneal, self-aligned $TiSi_2$ is formed to reduce the sheet resistance of gate and diffusion regions to 4–5 Ω/□. To minimize gate RC delay in high-speed circuits due to fine-line $TiSi_2$ resistance problems, a contact- and metal-over-gate scheme is implemented in ring oscillator and current-gain cut-off frequency (f_T) test sites [16].

• *Device characteristics*

Table 1 summarizes the measured 0.1-μm CMOS device parameters. Very low source–drain series resistances, 250 Ω-μm for n-MOSFETs and 700 Ω-μm for p-MOSFETs,

are obtained. Their effect on switching speed is minimal (≈10%). The abruptness of the extension profile is a key to achieving these low resistances.

The *I–V* characteristics of a 0.09-μm-channel n-MOSFET and an 0.11-μm-channel p-MOSFET are shown in **Figure 5**. The effective channel lengths are extracted from a series of low-drain-bias I_{DS}–V_G curves using the "shift and ratio" method [17]. The saturation transconductances are 620 mS/mm for n-MOSFETs, and 290 mS/mm for p-MOSFETs, respectively. The subthreshold characteristics of a 0.10-μm n-MOSFET and a 0.12-μm p-MOSFET are shown in **Figure 6**, where subthreshold slopes of 85–90 mV/decade and off-currents of less than 1 nA/μm are obtained. Significantly better short-channel effects are achieved with the halo implant [16]. The shortest devices obtained without punch-through (at V_{DS} = 1.5 V) are 0.091 μm for n-MOSFETs and 0.084 μm for p-MOSFETs. Even with the halo improvement, the experimentally measured short-channel effect and punch-through are significantly worse than those designed from 2D device models. There are many possible contributors to the discrepancy: boron depletion in short-channel devices, polysilicon gate etch profile, source–drain lateral gradient, and physical interpretation of extracted channel length. This is an area that clearly needs further work in order to improve the tolerances in a practical 0.1-μm CMOS technology.

The ac performance of 0.1-μm CMOS devices was evaluated using both f_T and ring oscillator test sites. The highest f_T obtained are 118 GHz for n-MOSFETs and 67 GHz for p-MOSFETs [11]. The gate delay of a 101-stage unloaded CMOS-inverter ring oscillator is shown in **Figure 7** as a function of power supply voltage. At 1.5 V, the delay is 22 ps/stage. This is more than a factor of 2 faster than the previous 0.25-μm CMOS devices operated at 2.5 V (solid square in Figure 7) [5]. The measured delays agree well with model simulations in Figure 7. There is a slight difference because of the inexact match of threshold voltage and channel length.

For low-power operation at lower supply voltages, the delay increases, but is still less than that of 0.25-μm CMOS, even below 1 V. Reasonably high G_ms (340 mS/mm for n-MOSFETs and 140 mS/mm for p-MOSFETs) are obtained at $V_{DS} = V_G$ = 0.6 V. Compared with the G_ms at 1.5 V, these values are lower by less than the $(V_{DD} - V_t)$ ratio, since high-field effects like velocity saturation and mobility degradation are not as severe [6]. Device reliability also improves significantly at such low operating voltages. At a 0.5-V power supply voltage, the delay is 95 ps per stage.

Power per stage of the 0.1-μm CMOS ring oscillator is plotted versus gate delay in **Figure 8**, where corresponding figures for 0.25-μm and 0.5-μm CMOS circuits are also shown for comparison. At the highest performance (20 ps), the power is not too much lower than for 0.25-μm CMOS.

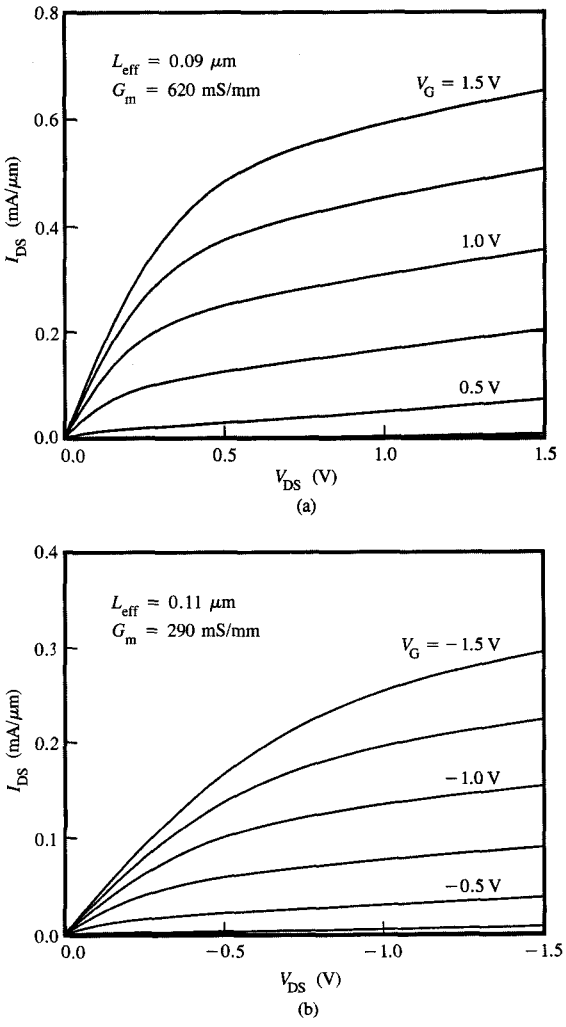

Figure 5

Measured *I–V* characteristics of (a) 0.09-μm n-MOSFET and (b) 0.11-μm p-MOSFET.

On the other hand, at the same delay as 0.25-μm CMOS, the power per stage of 0.1-μm CMOS is 21 times smaller. The power–delay product per stage of the 0.1-μm CMOS ring oscillator basically follows a CV^2 dependence (with $C \approx 25$ fF) as expected. Deviations occur below $V_{DD} \approx 0.6$ V when the standby power becomes appreciable. An ultralow (power per stage)–(delay per stage) product, 0.03 fJ per stage (switching factor = 0.01), with a gate delay of 190 ps is obtained at a power supply voltage of 0.4 V [6]. This corresponds to a switching energy of ≈2 fJ per transition for the 0.1-μm CMOS inverter (W_n = 3 μm, W_p = 4 μm).

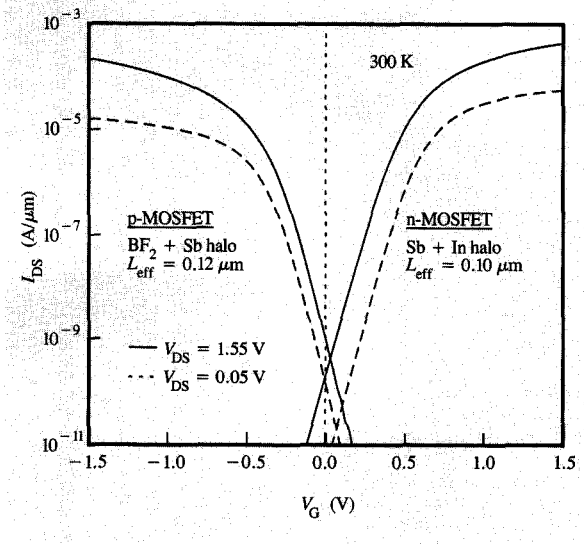

Figure 6

Measured subthreshold characteristics of 0.10-μm n-MOSFET and 0.12-μm p-MOSFET.

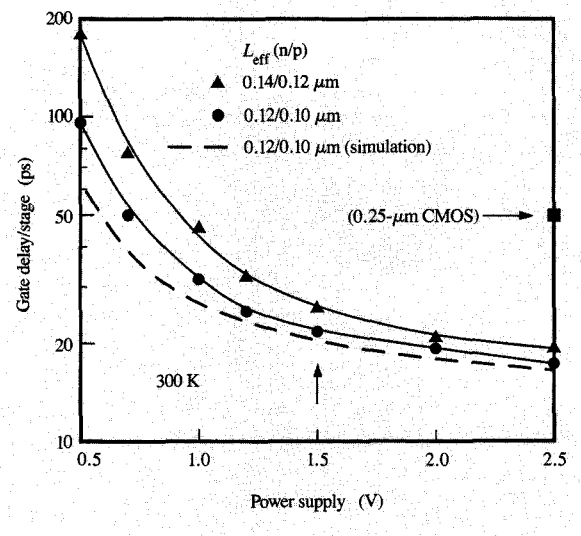

Figure 7

Measured (points, solid lines) and simulated (dashed line) CMOS-inverter delay vs. supply voltage.

3. Limit of scaling

Beyond 0.1-μm CMOS, a number of fundamental factors may impose a limit on device scaling. These are oxide and silicon tunneling, random dopant distribution, threshold voltage nonscaling, and interconnect delays. They are examined in the order listed.

• *Oxide tunneling*

Pushing CMOS scaling beyond the 0.1-μm channel length requires the use of ultrathin gate oxides with thicknesses less than 30 Å. As the oxide film becomes thinner, the gate leakage current, because of increasing direct quantum-mechanical tunneling, becomes significant. In **Figure 9**, the current density is shown for films in the 25–36-Å thickness regime from a number of different laboratories [18, 19]. Also shown is the current voltage relationship for a very thin (~18 Å) oxide with an aluminum gate. From these data, the effect of tunneling current on standby power consumption can be estimated. For example, for the 25-Å film, the current density with a 1-V supply is ~0.1 A/cm^2. Given that the total gate area on current and future ULSI logic chips will be of the order of 0.1 cm^2 or less, the power resulting from the gate leakage will be only about 10 mW. If the gate dielectric thickness were to decrease to 20 Å, the current density would increase to 1–10 A/cm^2, which in turn would increase the power consumption to 0.1–1 W. This would still be acceptable for high-performance logic chips whose power is normally in the 5–50-W range. The effect of gate tunneling on individual device operation should be negligible, since the per-width leakage current, 1 nA/μm for a tunneling current density of 1 A/cm^2, is many orders of magnitude below the device current at threshold, ~10 μA/μm. It appears, therefore, that the tunneling current in itself will not be a limiting factor, at least in terms of the standby power of logic chips, even for gate dielectric thicknesses in the 20–25-Å regime.

However, as the gate dielectric layer becomes thinner, device yield and/or reliability may become an issue. The energy of the electrons passing through the gate dielectric decreases substantially as the supply voltage of CMOS devices is scaled, but the electron fluence (i.e., the integrated electron flux through the gate oxide) increases, since the gate leakage current increases exponentially with decreasing dielectric thickness. In the ≤40-Å thickness regime, the electron transport in oxide is more or less ballistic, and the electron energy is governed by the applied bias [20]. For devices with gate dielectrics in this regime, the voltage drop across the oxide will be no more than 1.5 V, and the electron transport will be limited to direct quantum-mechanical tunneling. The electron energies, as determined by the maximum oxide voltage drop, will be low enough that oxide degradation and ultimately breakdown should be extended to much higher electron fluences. In other words, the defect generation rate should decrease drastically, which extends the dielectric reliability.

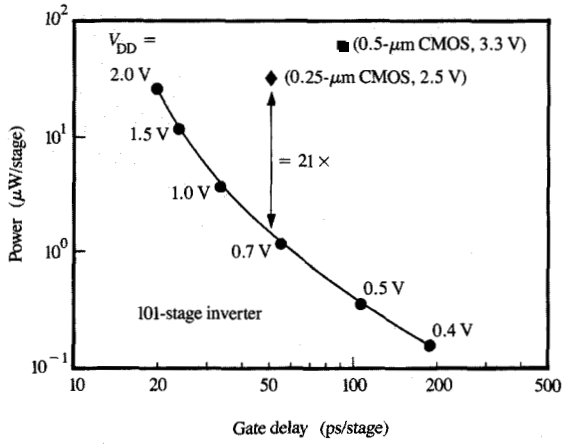

Figure 8

Measured power per stage vs. gate delay of 0.1-μm CMOS ring oscillator with V_{DD} as a parameter. The device widths are W_n = 3 μm, W_p = 4 μm.

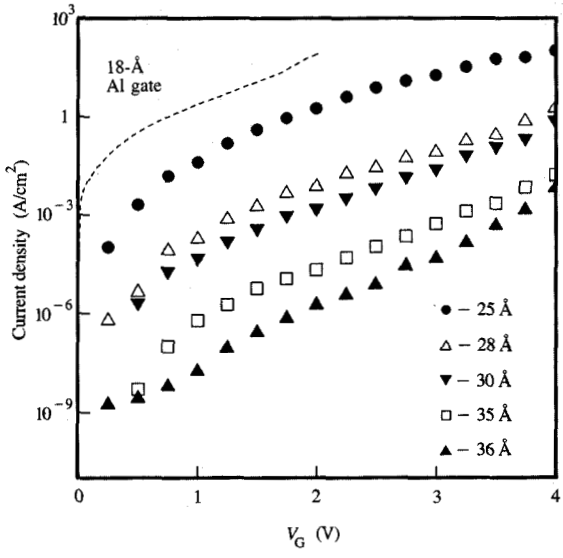

Figure 9

Current density–voltage relationships for polysilicon gate capacitors with gate oxide thicknesses ranging from 25–36 Å from a number of different research establishments [19, 20]. The dashed line is for an Al gate with an oxide ~18 Å thick.

Apart from the issue of oxide reliability, the question of yield may in fact be what ultimately determines the practical limit of such thin dielectrics. To manufacture a 20-Å gate oxide with ±10% (±2-Å) uniformity across a 200-mm wafer, with projected defect densities as low as 10^{-2} cm^{-2}, is, needless to say, a formidable task.

• *Silicon tunneling*

To control short-channel effect and prevent device punch-through, very high channel doping (1–5 × 10^{18}/cm^3) will be required for channel lengths of less than 0.1 μm. Such high doping concentrations, however, could cause significant tunneling current in source–drain junctions. For a silicon p–n junction, the expression for indirect tunneling current density [21] is

$$J_t = [(2m^*)^{1/2}q^3E_JV/(4\pi^3\hbar^2E_G^{1/2})]\exp[-4(2m^*)^{1/2}E_G^{3/2}/(3qE_J\hbar)], \quad (3)$$

where E_J is the maximum electric field in the junction, V is the applied bias, E_G is the bandgap, and m^* is the reduced effective mass of an electron. According to [21], an effective mass $m^* = 0.165m_0$ seems to fit the hardware data best, and is thus used here. The maximum electric field of a p–n junction can be determined using the following equation, valid for a one-sided abrupt junction (worst case):

$$E_J = [2qN_d(V + V_{bi})/\varepsilon_{Si}]^{1/2}. \quad (4)$$

A built-in voltage, V_{bi}, of 1 V is used here, following previous publications [22].

The calculated tunneling current density as a function of applied voltage is plotted in **Figure 10** for various substrate doping concentrations, N_d. A tunneling current limit of 1 A/cm^2 (dashed line) is depicted in the figure for comparison. This current corresponds to a junction leakage of 2–3 nA/μm for 0.1-μm devices, which is still less than the worst-case leakage current specification for high-performance logic circuits. Therefore, it appears that a substrate doping of up to 5 × 10^{18}/cm^3 can be used without significant increase in off-current. According to generalized scaling principles [23], such a doping concentration should provide a reasonable design point for channel lengths down to 0.05 μm, if a 20–25-Å-thick gate oxide is assumed.

• *Dopant fluctuations*

It was predicted in the 1970s [24, 25] that random fluctuation of the number of dopant atoms in the channel of a MOSFET would be a fundamental physical limitation of MOSFET miniaturization. As MOSFET scaling approaches the sub-0.1-μm regime, the number of dopants is of the order of hundreds in the depletion region, and less than 100 in the inversion layer, for minimum-geometry devices. As a result, the detailed microscopic dopant

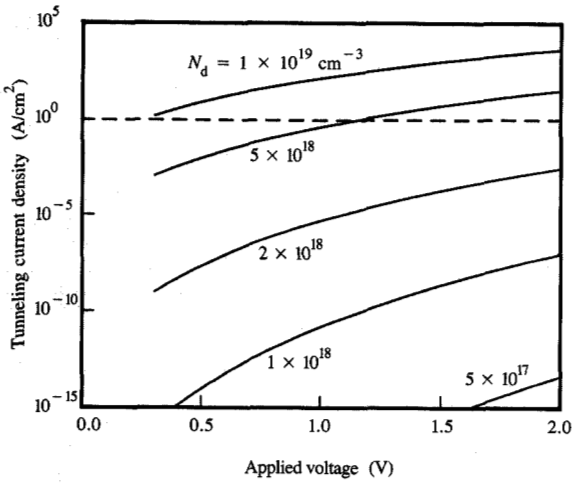

Figure 10

Calculated silicon tunneling current density vs. bias voltage for different background doping concentrations, N_d. The dashed line (1 A/cm^2) indicates the limit beyond which the standby power due to tunneling leakage could be appreciable.

distribution in the MOSFET channel will have nonnegligible influence on device electrical performance.

The dependence of the terminal currents and the threshold voltage on 1) the random fluctuation of the number of dopants in the MOSFET channel and 2) the discrete microscopic random distribution (arrangement) of dopant atoms in the MOSFET channel [26] was studied using a three-dimensional drift-diffusion simulation program, FIELDAY [27]. **Figure 11** shows a sample set of I–V curves of 24 MOSFETs with different random dopant "atom" distributions. When compared with the I–V curve of the same MOSFET simulated using the conventional continuum doping model, the discrete doping simulation displayed 1) a spread of the I–V curves along the gate voltage axis of about 20 mV (one standard deviation); 2) an average shift of the I–V toward the negative gate voltage direction of about 30 mV in the subthreshold region and of about 15 mV in the linear region; and 3) a slight degradation (<3 mV/decade) and fluctuation of the subthreshold slope. The V_t shift in the subthreshold region was obtained by current averaging (the triangular curve in Figure 11) and was larger than in the linear region because of the logarithmic dependence. Furthermore, the I–V curves of narrow-width devices were asymmetric upon interchanging the source and the drain terminals. The asymmetry of threshold could be as large as 60 mV. This asymmetry is attributed to the discrete nature of the dopant atoms, which resulted in an inhomogeneous channel potential.

The effects of discrete random dopants become more important as the gate-controlled channel volume decreases (e.g., decreasing channel length or increasing drain voltage). Assuming an n+-polysilicon gate for n-MOSFETs, and vice versa, the fractional threshold voltage uncertainty due to random dopant fluctuations [28] can be shown analytically to be

$$\sigma_{V_t}/V_t \approx 1/(3N_a W_m L W)^{1/2}, \tag{5}$$

where σ_{V_t} is the standard deviation of V_t fluctuation, N_a is the background doping concentration, W_m is the maximum depletion width, and L and W are the channel length and width. For a MOSFET with $W = L = 0.05$ μm and $N_a \approx 5 \times 10^{18}$/cm^3, $W_m \approx 150$ Å and $\sigma_{V_t}/V_t \approx 4\%$. This is still manageable (less than the threshold voltage variation allotted for the short-channel effect) and does not impose a fundamental limit to miniaturization at the 0.05-μm-channel generation.

The major impact of discrete random dopants on device miniaturization is likely to be in two areas: off-current estimation and modeling, and threshold voltage control and matching. The threshold voltage shift in the subthreshold region means that conventional estimations of off-current (hence standby power) from the linear threshold voltage could be about a factor of 2 too low. Threshold matching is particularly important for certain types of circuits such as SRAM, where minimum-geometry devices are often employed. For bulk as well as silicon-on-insulator (SOI) MOSFETs, the channel volume continues to decrease as devices are miniaturized. Ideally, one can circumvent the dopant fluctuation problem altogether by using a very thin, undoped channel and controlling threshold voltage by the gate work function. This is discussed further in connection with the double-gate MOSFET described in Section 4.

- *Subthreshold leakage and standby power*

Among the roadblocks to MOSFET miniaturization, subthreshold nonscaling is the most serious threat to continuing performance improvement. The subthreshold slope, of the order of $(\ln 10)(kT/q)$, is independent of oxide thickness, channel length, and supply voltage. To keep off-current within standard specifications, the threshold voltage cannot be reduced appreciably, as indicated in Figure 1. CMOS logic technologies with channel lengths of 0.25 μm and less must deal with this issue and often must trade off performance for lower off-current.

Keeping a constant power supply voltage, say >2 V, as the channel length is scaled down is not an acceptable solution, as can be seen from the active power equation, (1). For future high-performance microprocessor chips

using more advanced lithography, both C_{sw} and f will increase to provide faster computational capabilities through higher integration and increased clock speeds. This is evidenced by the fact that the die size, instead of shrinking by the square of the scaling factor for the same circuit count, has actually increased slightly with a rapidly growing number of circuit counts over the generations [29]. The trend is expected to continue, since more microprocessor performance gain is likely to come from parallelism in the future. The only way, then, to keep active chip power in manageable limits without an expensive cooling package is to reduce the supply voltage, V_{DD}. If V_t is scaled down in step with V_{DD}, it will lead to subthreshold MOSFET currents in standby which grow exponentially with decreasing lithography scale. This passive power component is particularly troublesome for low-power or portable systems, since traditional power management systems are unable to circumvent such currents; the subthreshold currents are present whether or not the circuits are in operation.

On the other hand, if one keeps V_t at, say, 0.4–0.5 V, to constrain the passive power while V_{DD} is reduced, the MOSFET performance in ULSI designs will reach saturation in the 0.1-to-0.2-μm lithography scale unless other power-avoidance techniques are utilized [29].

While architectural innovations are likely to continue to help allow active power reduction, and thus allow higher V_{DD} for a given lithography scale, CMOS technology is likely also to experience some pressures for change to assist in power reduction. A widely discussed feature is the addition of a second, higher-V_t MOSFET in future technologies. Low-V_t MOSFETs may be used as traditional (CMOS) circuit elements, while a single high-V_t "footswitch" would source the circuit's ground current. This may be the ultimate solution to the standby power and V_t problem. One can use low-V_t devices in critical logic paths for speed while using high-V_t devices everywhere else (including the memory) to minimize standby power. One can also sense the circuit activity and cut off the supply to logic devices that are not switching. The process can be done simply by adding a couple of block-out masks. Many circuit schemes, such as "domino logic," are ideally suited for such an approach. This would allow V_t reduction in step with scaling, while managing subthreshold currents for the entire die to an acceptable level. Low body-effect pass gates could also be made available through this means to avoid the increasingly difficult performance issues associated with scaled-V_{DD} latches. However, dealing with noise margin and inductive effects might make such approaches more difficult to implement in the highest-performance systems. Threshold-voltage engineering is likely to become a central issue in sub-0.1-μm CMOS technologies.

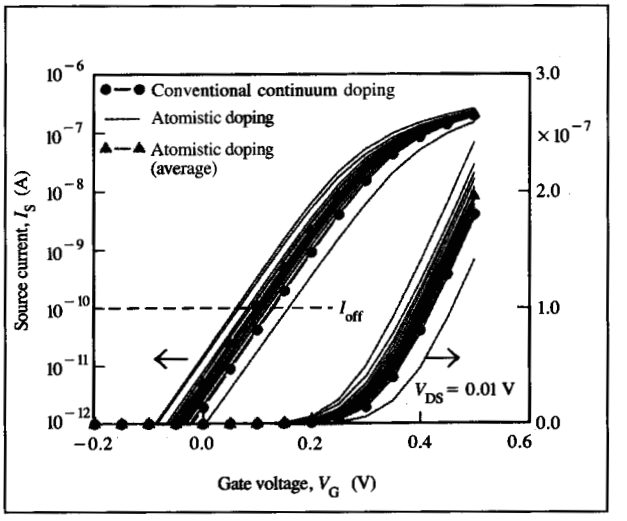

Figure 11

Simulated source current vs. gate voltage for a conventionally doped MOSFET (solid dots) and 24 devices with different discrete dopant distributions in the channel (grey lines). Channel length is 0.1 μm; channel width is 0.05 μm. Average current of all 24 devices is shown in solid triangles. Threshold voltage shift in the subthreshold region is defined as the gate voltage shift at which the source current is equal to I_{off}.

Interconnect delays

The discussions so far have been focused primarily on devices. At the chip level, however, a key issue that must be addressed is the interconnect RC delay, which quickly becomes a serious problem as the lithography scale shrinks and the clock frequency increases.

Delays stemming from wire resistance, to first order, do not decrease in spite of scaling to smaller dimensions. The factor that improves the wire RC delay [30],

$$\tau_w = 0.5 R_w C_w W_L^2, \tag{6}$$

through shorter wire length, W_L, is negated by the increase in wire resistance per unit length, R_w, due to wire cross-section shrinkage. Wire capacitance per unit length, C_w, in the meantime remains constant, around 0.2 pF/mm for minimum-width wires with oxide dielectric. If one takes a 1-μm × 1-μm-cross-section, typical back-end-of-line (BEOL) aluminum metallurgy wire with oxide dielectric, the wire itself introduces ~100 ps delay if it runs 4 mm, and ~900 ps for 12 mm. Such RC wire delay values will remain characteristic of half- and full-chip-length wires, respectively, unless something quite different is done. As long as one is dealing with cycle times greater than 3–4 ns, the wire RC delays are barely noticeable. However, for ultrahigh-performance CPUs, with cycle times around and below 2 ns, the resistance can be a make-or-break

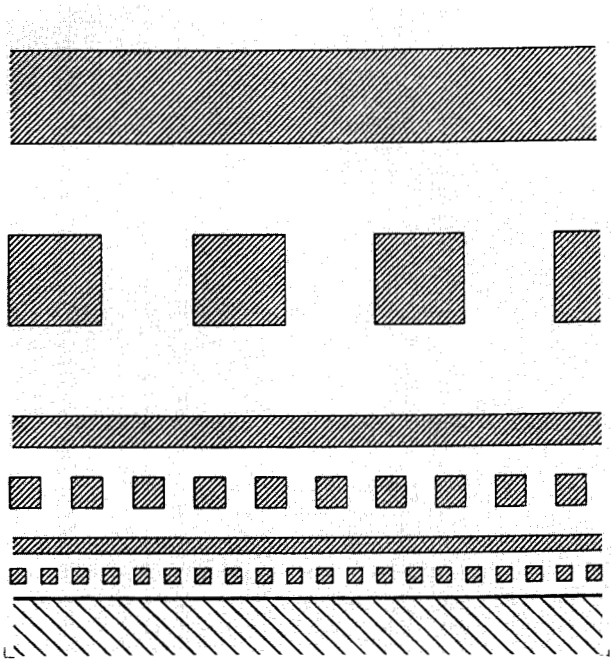

Figure 12

Schematic cross section of levels of desired chip wiring for high-performance CMOS processors.

proposition for CMOS processors. The use of repeaters, regenerating the signal along the way, is helpful, since it decreases the dependence on the wire length from square to a linear one. But repeaters alone do not solve the problem, since delays still remain in an unacceptable range and they introduce additional problems in increased power consumption and design complexity.

If one looks carefully at the roles various interconnections play, a better approach suggests itself [31, 32]. High-performance processors require two kinds of wires. First, there are the wires that serve for the vast majority of interconnections. For CMOS processors, these "short" wires are typically at most a few hundred microns long. They make the chip "wirable" by providing a sufficient number of interconnections. Here the *RC* component plays no appreciable role. Such "short" wires should scale proportionally as lithography becomes smaller. Second, there is a need for "long" wires, where density is secondary to delay considerations. These interconnections earlier were part of the package, but with integration they are now on the chip. They run between distant parts of the chip, and their characteristic length is that of a chip-edge. A good scaling gauge for such "long" wires is that the time of signal propagation on them should be a small fraction of the processor cycle time. From such considerations, it immediately follows that these wires cannot be reduced in the same proportion as other features. On the contrary, with decreasing cycle times their size, pitch, and interlevel separation may actually have to increase. These will be referred to as "fat" wires. **Figure 12** shows in cross section an example of the interconnection scheme needed by ultrahigh-performance processors. It features a hierarchy of three x–y wiring-level pairs. The levels on the bottom are at the finest pitch of which the given technology is capable. The next two levels already pay attention to the *RC* problem, and finally the top two can serve to run signals to full chip-edge-length, or longer, distances. With this type of wiring, where conductor and dielectric cross-sectional dimensions are scaled together, capacitance per unit length stays constant for each level, while resistance decreases proportionally with wire cross-section increase. In Figure 12, $R \times C$ in the second x–y plane is a fourth, and in the third a 36th, of that of the bottom plane.

One consequence of having low-*RC* wires is that one observes transmission-line characteristics not only on the package, but also on the chips themselves. When the input of a wire is driven with a faster signal than the travel time down that line, delays are necessarily dominated by transmission-line characteristics, and finite signal-propagation speed must be taken into account. With an oxide insulator, the minimum delay that a signal can achieve due to the finite velocity of electromagnetic wave propagation is ~7 ps/mm. For example, on a 15-mm-long wire, the signal flight time cannot be less than 105 ps. This is significantly longer than the switching time of drivers in the considered technologies.

Figure 12 serves as illustration only; the number of wiring planes and the ratios between them must be optimized for any given design. However, for reaching the highest performances, the ratios shown are quite realistic. There are no possible materials, neither metals nor insulators, which could give the needed low-*RC* delays without the "fat" wire scheme. Accordingly, for CMOS processors there is a split between the needs of systems that stress cost and/or low power, and high-performance systems. The wiring presented in Figure 12 serves the purposes of performance-oriented processors. Larger system area and higher power are the penalties associated with it.

The net result is that with the proper kind of wiring one can avoid a so-called "*RC* crisis." The scheme described above reduces the problem to one of coping with time-of-flight delays, which, for CMOS at least, is a much less severe restriction on performance.

4. Novel devices beyond 0.1 μm

To go beyond 0.1-μm CMOS, that is, to exceed the minimum threshold and power supply voltage limits mentioned in Section 2, is difficult at room temperature for

conventional circuits, unless one is willing to either relax the off-current requirement or forgo the performance gain. There are, however, a few material- and/or structure-related device possibilities for performance improvement beyond 0.1 μm. These are SiGe channel, SOI, low-temperature CMOS, and double-gate MOSFET, as discussed below.

• *SiGe and SOI devices*

Recently, SiGe-channel p-MOSFETs have been fabricated using a UHV–CVD process [33]. Up to 50% higher hole mobilities, attributed to lighter effective mass and valence-band offset, have been reported. However, it is not clear whether SiGe can offer similar improvement in the n-channel MOSFET, which is the more important device from a circuit-performance point of view. With the hole mobility improvement alone, only 10–15% higher CMOS performance can be expected. There are also substantial integration issues, since p-MOS improvement should not be achieved at the expense of n-MOS. Furthermore, gains in short-channel devices due to higher mobility in SiGe are limited, since the saturation velocity remains basically the same as that of silicon [33].

Another way to enhance CMOS performance is to use SOI substrates. The main advantage stems not from dc currents but from reduced parasitic (diffusion and substrate) capacitances. A factor of 1.3–2.0 improvement in CMOS circuit speed has been reported with SOI devices [34]. Besides SOI material and cost issues, however, there are undesirable floating-body effects which cause a strong V_t dependence on drain voltage due to impact ionization at the drain end of the channel. This tends to limit SOI devices to either low power-supply voltages or fully depleted operation. Fully depleted operation would require very thin SOI films, which could have source–drain contact resistance problems [15]. A possible solution is to use selective epitaxial deposition to form a raised source–drain region for contacts. More extensive discussion of SOI devices can be found in a separate paper in this issue [35].

• *Low-temperature CMOS*

For high-performance systems, low-temperature-operated CMOS is also a possibility. The performance advantages of low-temperature FET operation have been recognized and advocated for a long time [36]. It appears, however, that as long as performance improvements can be made at room temperature, low-temperature operation will remain a matter of discussion only. Since now we are perceiving limits in room-temperature CMOS performance, we must begin to take low-temperature CMOS seriously.

Because of higher carrier mobility and lower interconnect resistance, low-temperature CMOS can provide a factor of 2 performance gain over room-temperature CMOS [37]. More importantly, at low

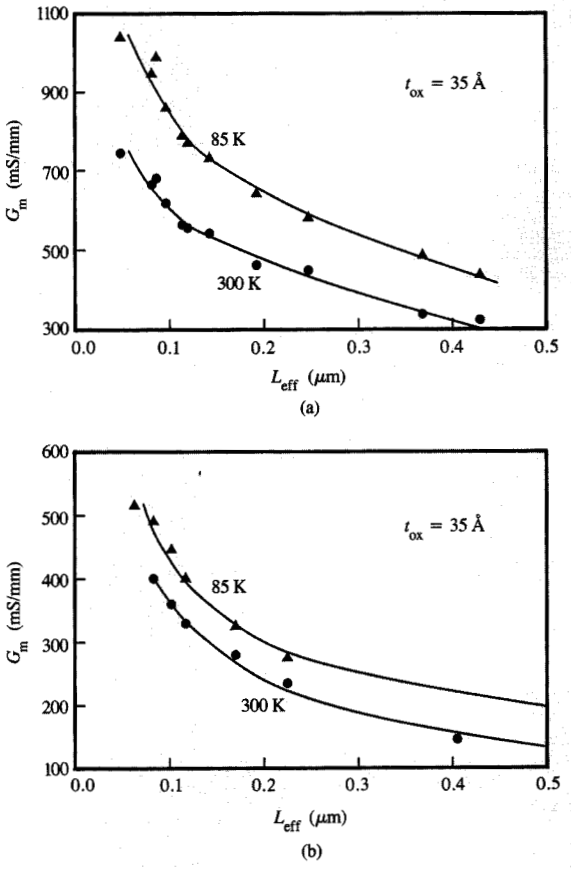

Figure 13

Measured (a) n-MOSFET and (b) p-MOSFET saturation transconductances at 300 and 85 K vs. channel length.

temperature the channel can be shortened further, with continuing performance gains. The fundamental reason for the scalability of FETs at low temperature is that devices can be turned off much more readily than at room temperature. This fact allows for a whole different, low-threshold, low-voltage design space from which room-temperature operation is excluded [2, 23] (unless the multiple-threshold voltage scheme discussed in the subsection on subthreshold leakage and standby power can be implemented). However, steeper subthreshold slope by itself is not sufficient for operating at a low threshold voltage. Very tight threshold tolerance is required as well, which will be a key challenge for low-temperature CMOS.

In **Figure 13**, the measured room- and low-temperature saturation transconductances are plotted versus channel length, where values of 1040 mS/mm for an n-MOSFET [16] and 510 mS/mm for a p-MOSFET [14] at low

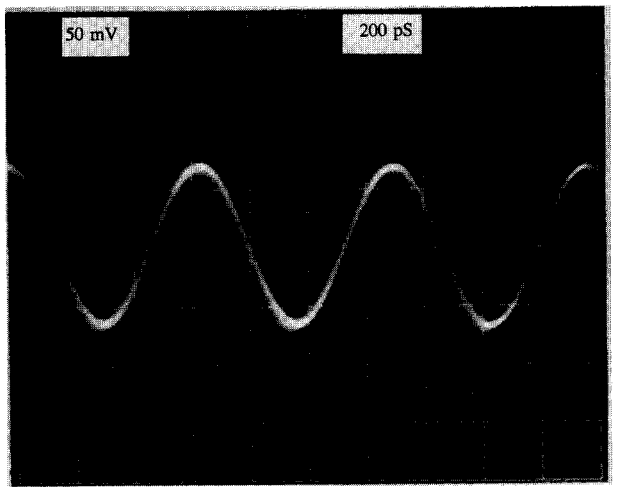

Figure 14

Measured waveform of a 43-stage, 0.08-μm-channel n-MOS ring oscillator at 85 K. Gate delay is 7.8 ps per stage.

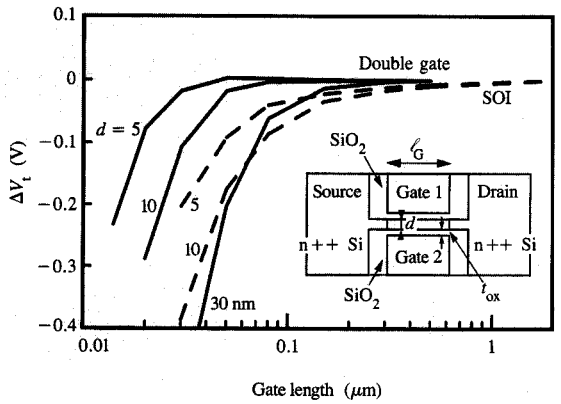

Figure 15

Simulated threshold voltage vs. channel length, comparing short-channel effect of double-gated FETs (solid lines) with SOI MOSFETs (dashed lines), where the threshold of the long-channel FETs has been taken as zero. These values are extracted from drift diffusion simulations of the subthreshold regime of these FETs. Inset: Cross-sectional structure of a double-gated MOSFET.

temperatures are the highest reported to date. **Figure 14** shows the waveform of 43-stage inverter-type n-MOS ring oscillators at 85 K. A minimum delay of 7.8 ps per stage is obtained from the 0.08-μm channel ring oscillator operating at 2.5 V [16]. This is the fastest switching speed reported to date for any silicon device at any temperature.

• *Monte Carlo simulation of a 30-nm double-gated MOSFET*

In an effort to understand the outermost limits of scaling, recent simulation studies have focused on double-gated FETs [38, 39]. The structure of these FETs is sketched in the inset to **Figure 15**. There is a very thin Si layer for a channel, with two gates, one on each side of the channel. The two gates are electrically connected together so that they both serve to modulate the channel. Short-channel effects are greatly suppressed in such a structure because the two gates very effectively terminate the drain field lines, preventing the drain potential from being felt at the source end of the channel. Consequently, the variation of the threshold with drain voltage and with gate length of a double-gated FET is much smaller than that of a conventional single-gated structure of the same channel length. This can be seen in Figure 15, where the threshold-versus-gate-length behavior of the double-gated MOSFET is compared with that of single-gated SOI MOSFETs. Note that for the same channel thickness, the double-gated FETs can be scaled to 2–3 times shorter channel lengths.

To estimate a limit on the scaling of such double-gated FETs, it is necessary to consider various device physics principles and tolerance issues. Since voltages must be low, the threshold voltage uncertainty should be kept to 100 mV or less. Channel thickness uncertainty causes uncertainty in the energy of the first quantized energy level of the channel, which translates into threshold voltage variation. This uncertainty grows very rapidly as the channel is thinned, which results in a minimum viable channel thickness of 4–5 nm, assuming a thickness tolerance of ~20%. Given a 5-nm-thick channel and 3-nm-thick gate oxide, Figure 15 indicates a minimum channel length of 30 nm using the criterion of 100-mV threshold variation for a 30% gate-length variation. To avoid threshold fluctuations due to the discreteness of the dopants, it would be necessary to adjust the threshold of this FET by the workfunction of the gate, leaving the channel undoped.

To evaluate the potential on-state performance of these FETs, detailed Monte Carlo simulations have been performed [38, 39] using the simulator DAMOCLES [40]. Both n- and p-channel MOSFETs have been simulated, yielding low-output-conductance, high-performance I–V characteristics for both device types, as is illustrated in **Figure 16** for the n-FET. The transconductance exceeds 2300 mS/mm for this n-FET, and it reaches 1300 mS/mm for the p-FET. Transient Monte Carlo simulations have also been done for an n-FET switching a capacitive load equivalent to another n-FET. This resulted in a minimum estimated switching time of 1.1 ps for this n-FET, clearly indicating the potential for performance in these tiny FETs.

The Monte Carlo simulations also allow an analysis of the internal carrier behavior of the double-gate MOSFET.

As illustrated in **Figure 17**, the carriers behave quite ballistically in these short devices. Very little kinetic energy is lost until the carriers reach the drain end of the device. In keeping with this observation, the electrons reach peak velocities as high as 3×10^7 cm/s just before entering the drain. The holes, however, only reach 1.3×10^7 cm/s, even though they lose relatively little energy. It appears that a high-momentum scattering rate is responsible for reducing the hole velocities and currents to only about half those of the electrons in the n-FET, even at the limits of scaling.

5. Conclusion

In conclusion, high-performance 0.1-μm CMOS devices operating at a 1.5-V power supply voltage have been demonstrated. Key technology features include dual n+/p+ polysilicon gates on 35-Å gate oxide, retrograde channel profile, and 500-Å-deep source–drain extensions with self-aligned halo. High f_T values, 118 GHz for n-MOSFETs and 67 GHz for p-MOSFETs, have been obtained. A 22-ps-per-stage unloaded CMOS-inverter ring-oscillator delay is achieved at a 1.5-V power supply voltage, which represents a factor of 2 performance gain over the 0.25-μm CMOS technology at 2.5 V. In addition, ultralow-power 0.1-μm CMOS is demonstrated at power supply voltages well below 1 V. A 20× reduction in active power per circuit compared with the 0.25-μm CMOS is obtained at the same delay as the 0.25-μm CMOS. An ultralow switching energy, 2 fJ per transition, is achieved at a 0.4-V supply voltage. These results clearly establish the feasibility of 0.1-μm CMOS for both high-performance and low-power VLSI applications.

A number of key device and technology issues which may ultimately determine the limit of room-temperature scaling have been examined. It is felt that, although a great deal of effort is needed to overcome these problems, oxide and silicon tunneling, dopant fluctuations, and interconnect *RC* delays do not impose a fundamental limit on CMOS scaling to 0.05-μm channel length. Off-current leakage due to subthreshold nonscaling, however, is a more serious problem and may require circuit solutions. A promising approach would be to fabricate multiple threshold-voltage devices on a chip to manage standby power without degrading performance. On alternative material and device structures, limited performance enhancement can be obtained with SiGe channel and SOI devices without channel-length shrinkage. Low-temperature CMOS and double-gate MOSFETs, on the other hand, can not only provide a factor of 2 performance gain but also extend channel-length scaling to the shortest possible limit. The challenges, however, lie in the fabrication of double-gate MOSFETs and low-cost cooling of VLSI chips/packages in a room-temperature environment.

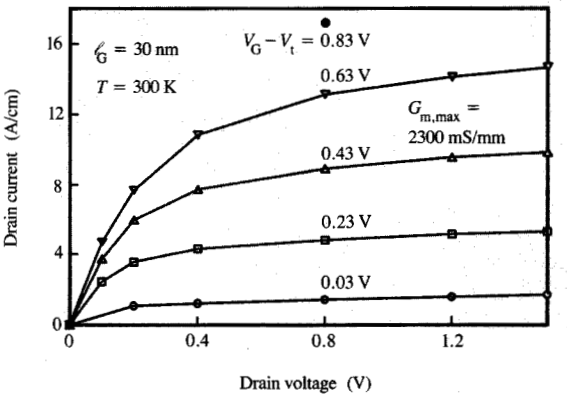

Figure 16

Monte Carlo simulation of drain current vs. drain voltage for an n-channel double-gated MOSFET. Channel length is 30 nm; channel thickness is 5 nm. Note the high transconductance (2300 mS/mm) and the low output conductance.

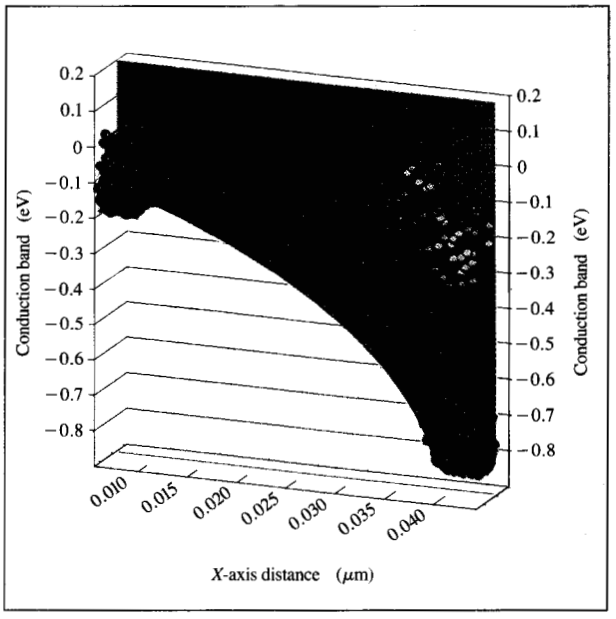

Figure 17

Monte Carlo simulation of electron energy vs. position down the channel of an n-channel double-gated MOSFET. The points represent electrons, and the line indicates the conduction band edge. The height of the points above the band edge indicates their kinetic energy.

Acknowledgments

The authors would like to thank D. Kern, T. H. Ning, M. R. Wordeman, G. G. Shahidi, and R. H. Dennard for many helpful discussions. They would also like to thank M. R. Polcari, T. P. Smith III, and P. M. Horn for management support; and the Yorktown Silicon Facility and E-Beam Group for device fabrication.

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 Circuits* **SC-9,** No. 5, 256 (1974).
2. G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, S. A. Rishton, E. Ganin, T. H. P. Chang, and R. H. Dennard, "Experimental Technology and Performance of 0.1-μm-Gate-Length FETs Operated at Liquid-Nitrogen Temperature," *IBM J. Res. Develop.* **34,** No. 4, 452 (1990).
3. F. S. Lai, L. K. Wang, Y. Taur, Y. C. Sun, K. E. Petrillo, S. M. Chicotka, E. J. Petrillo, M. R. Polcari, T. J. Bucelot, and D. S. Zicherman, "A Highly Latchup-Immune 1-μm CMOS Technology Fabricated with 1-MeV Ion Implantation and Self-Aligned $TiSi_2$," *IEEE Trans. Electron Devices* **ED-33,** No. 9, 1308 (1986).
4. L. K. Wang, Y. Taur, D. Moy, R. H. Dennard, F. Hohn, P. J. Coane, A. Edenfeld, S. Carbaugh, D. Kenny, and S. Schnur, "0.5-μm Gate CMOS Technology Using E-Beam/Optical Mix Lithography," *1986 IEEE Symposium on VLSI Technology Digest of Technical Papers*, p. 13 (1986).
5. W. H. Chang, B. Davari, M. R. Wordeman, Y. Taur, C. C. H. Hsu, and M. D. Rodriguez, "A High Performance 0.25 μm CMOS Technology: I-Design and Characterization," *IEEE Trans. Electron Devices* **ED-39,** No. 4, 959 (1992).
6. Y. Mii, S. Wind, Y. Taur, Y. Lii, D. Klaus, and J. Bucchignano, "An Ultra-Low Power 0.1 μm CMOS," *1994 IEEE Symposium on VLSI Technology Digest of Technical Papers*, p. 9 (1994).
7. J. Maserjian and G. P. Petersson, "Tunneling Through Thin MOS Structures: Dependence on Energy (E-k)," *Appl. Phys. Lett.* **25,** No. 1, 50 (1974).
8. M. Endo, K. Hashimoto, K. Yamashita, A. Katsuyama, T. Matsuo, Y. Tani, M. Sasago, and N. Nomura, "Challenges in Excimer Laser Lithography for 256M DRAM and Beyond," *IEDM Tech. Digest*, p. 45 (1992).
9. J. Warlaumont, "X-Ray Lithography: On the Path to Manufacturing," *J. Vac. Sci. Technol. B* **7,** No. 6, 1634 (1989).
10. M. R. Wordeman, "Design and Modeling of Miniaturized MOSFETs," Ph.D. Thesis, School of Engineering and Applied Science, Columbia University, New York, 1986.
11. Y. Taur, S. Wind, Y. J. Mii, T. Lii, D. Moy, K. A. Jenkins, C. L. Chen, P. J. Coane, D. Klaus, J. Bucchignano, M. Rosenfeld, M. G. R. Thompson, and M. Polcari, "High Performance 0.1μm CMOS Devices with 1.5V Power Supply," *IEDM Tech. Digest*, p. 127 (1993).
12. G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, E. Ganin, S. Rishton, D. S. Zicherman, H. Schmid, M. R. Polcari, H. Y. Ng, P. J. Restle, T. H. P. Chang, and R. H. Dennard, "Design and Experimental Technology for 0.1μm-Gate-Length Low-Temperature Operation FET's," *IEEE Electron Device Lett.* **EDL-8,** No. 10, 463 (1987).
13. M. A. Gesley, F. J. Hohn, R. G. Viswanathan, and A. D. Wilson, "A Vector-Scan Thermal-Field Emission Nanolithography System," *J. Vac. Sci. Technol. B* **6,** 2014 (1988).
14. Y. Taur, S. Cohen, S. Wind, T. Lii, C. Hsu, D. Quinlan, C. Chang, D. Buchanan, P. Agnello, Y. Mii, C. Reeves, A. Acovic, and V. Kesan, "High Transconductance 0.1 μm pMOSFET," *IEDM Tech. Digest*, p. 901 (1992).
15. Y. Taur Y. C. Sun, D. Moy, L. K. Wang, B. Davari, S. P. Klepner, and C. Y. Ting, "Source–Drain Contact Resistance in CMOS with Self-Aligned $TiSi_2$," *IEEE Trans. Electron Devices* **ED-34,** No. 3, 575 (1987).
16. Y. Mii, S. Rishton, Y. Taur, D. Kern, T. Lii, K. Lee, K. A. Jenkins, D. Quinlan, T. Brown, Jr., D. Danner, F. Sewell, and M. Polcari, "Experimental High Performance Sub-0.1 μm Channel nMOSFET's," *IEEE Electron Device Lett.* **EDL-15,** No. 1, 28 (1994).
17. Y. Taur, D. S. Zicherman, D. R. Lombardi, P. J. Restle, C. H. Hsu, H. I. Nanafi, M. R. Wordeman, B. Davari, and G. G. Shahidi, "A New Shift and Ratio Method for MOSFET Channel-Length Extraction," *IEEE Electron Device Lett.* **13,** No. 5, 267 (1992).
18. M. Depas, B. Vermeire, P. W. Mertens, M. Meuris, and M. M. Heyns, "Ultra-Thin Gate Oxide Yield and Reliability," *1994 IEEE Symposium on VLSI Technology Digest of Technical Papers*, p. 23 (1994).
19. K. F. Schuegraf and C. Hu, "Oxide Breakdown Model for Very Low Voltages," *1993 IEEE Symposium on VLSI Technology Digest of Technical Papers*, p. 43 (1993).
20. D. J. DiMaria, E. Cartier, and D. Arnold, "Impact Ionization, Trap Creation, Degradation, and Breakdown in Silicon Dioxide Films on Silicon," *J. Appl. Phys.* **73,** 3367 (1993).
21. R. B. Fair and H. W. Wivell, "Zener and Avalanche Breakdown in As-Implanted Low-Voltage Si n-p Junctions," *IEEE Trans. Electron Devices* **ED-23,** No. 5, 512 (1976).
22. A. G. Chynoweth, W. L. Feldmann, C. A. Lee, R. A. Logan, and G. L. Pearson, "Internal Field Emission at Narrow Silicon and Germanium p-n Junctions," *Phys. Rev.* **118,** No. 2, 425 (1960).
23. G. Baccarani, M. R. Wordeman, and R. H. Dennard, "Generalized Scaling Theory and Its Application to a 1/4 Micrometer MOSFET Design," *IEEE Trans. Electron Devices* **ED-31,** No. 4, 452 (1984).
24. B. Hoeneisen and C. A. Mead, "Fundamental Limitations in Microelectronics—I. MOS Technology," *Solid State Electron.* **15,** 819 (1972).
25. R. W. Keyes, "Effect of Randomness in the Distribution of Impurity Ions on FET Thresholds in Integrated Electronics," *IEEE J. Solid-State Circuits* **SC-10,** 245 (1975).
26. 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. Digest*, p. 705 (1993).
27. E. Buturla, J. Johnson, S. Furkay, and P. Cottrell, "A New 3-D Device Simulation Formulation," *NASCODE VI: Sixth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits*, Boole Press, Dublin, 1989, p. 291.
28. T. Mizuno, J. Okamura, and A. Toriumi, "Experimental Study of Threshold Voltage Fluctuations Using an 8K MOSFET Array," *1993 IEEE Symposium on VLSI Technology Digest of Technical Papers*, p. 41 (1993).
29. E. J. Nowak, "Ultimate CMOS ULSI Performance," *IEDM Tech. Digest*, p. 115 (1993).
30. H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley Publishing Co., Reading, MA, 1990, Ch. 5.
31. G. A. Sai-Halasz, "Directions in Future High-End Processors," *Proceedings of the IEEE International Conference on Computer Design*, 1992, p. 230.

32. G. A. Sai-Halasz, "Performance Trends in High-End Processors," *Proc. IEEE* **83,** 20 (1995).
33. S. Subbanna, V. P. Kesan, M. J. Tejwani, P. J. Restle, D. J. Mis, and S. S. Iyer, "Si/SiGe p-Channel MOSFETs," *1991 IEEE Symposium on VLSI Technology Digest of Technical Papers*, p. 103 (1991).
34. G. Shahidi, B. Davari, Y. Taur, J. Warnock, M. R. Wordeman, P. McFarland, S. Mader, M. Rodriguez, R. Assenza, G. Bronner, B. Ginsberg, T. Lii, M. Polcari, and T. H. Ning, "Fabrication of CMOS on Ultrathin SOI Obtained by Epitaxial Lateral Overgrowth and Chemical-Mechanical Polishing," *IEDM Tech. Digest*, p. 587 (1990).
35. G. G. Shahidi, J. D. Warnock, J. Comfort, S. Fischer, P. A. McFarland, A. Acovic, T. I. Chappell, B. A. Chappell, T. H. Ning, C. J. Anderson, R. H. Dennard, J. Y. Sun, M. R. Polcari, and B. Davari, "CMOS Scaling in the 0.1-μm, 1.X-Volt Regime for High-Performance Applications," *IBM J. Res. Develop.* **39,** No. 1/2, 229 (1995, this issue).
36. 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* **ED-24,** 218 (1977).
37. Y.-C. Sun, Y. Taur, R. H. Dennard, and S. P. Klepner, "Submicrometer-Channel CMOS for Low-Temperature Operation," *IEEE Trans. Electron Devices* **ED-34,** No. 1, 19 (1987).
38. D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte-Carlo Simulation of a 30 nm Dual-Gate MOSFET: How Short Can Si Go?" *IEDM Tech. Digest*, p. 553 (1992).
39. D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo Simulations of p- and n-Channel Dual-Gate MOSFETs at the Limits of Scaling," *IEEE Trans. Electron Devices* **40,** 2103 (1993).
40. S. E. Laux, M. V. Fischetti, and D. J. Frank, "Monte Carlo Analysis of Semiconductor Devices: The DAMOCLES Program," *IBM J. Res. Develop.* **34,** 466 (1990).

Received June 1, 1994; accepted for publication October 14, 1994

Yuan Taur *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (TAUR at YKTVMV).* Dr. Taur joined IBM Research at Yorktown Heights as a research staff member in 1981. He received his B.S. degree in physics from the National Taiwan University, Taipei, in 1967 and his Ph.D. degree in physics from the University of California, Berkeley, in 1974. From 1981 to the present, Dr. Taur has been with the Silicon Technology Department of the Thomas J. Watson Research Center. His research activities include latchup-free 1-μm CMOS, self-aligned $TiSi_2$, 0.5-μm CMOS and bi-CMOS, shallow-trench isolation, 0.25-μm CMOS with n+/p+ polysilicon gates, SOI, low-temperature CMOS, and 0.1-μm CMOS. He is currently manager of the Exploratory Devices and Processes group. Dr. Taur is a senior member of the IEEE. He has served on the technical program committees and as a panelist of the IEEE Device Research Conference, the International Electron Devices Meeting, and the Symposium on VLSI Technology. He has authored or co-authored more than ninety technical papers and holds three U.S. patents. Dr. Taur has received three IBM Outstanding Technical Achievement Awards and three IBM Invention Achievement Awards.

Yuh-Jier Mii *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (mii@watson.ibm.com).* Dr. Mii received his B.S. degree in electrical engineering from the National Taiwan University, Taipei, in 1982, and his Ph.D. degree in electrical engineering from the University of California at Los Angeles in 1990. His thesis work included the introduction and development of infrared modulators using the strongly enhanced Stark effect of intersubband transitions in asymmetric quantum wells, and the investigation of hole intersubband absorption in SiGe/Si quantum wells. From 1990 to 1991, Dr. Mii worked at AT&T Bell Laboratories, Murray Hill, New Jersey, on high-electron-mobility SiGe/Si heterostructures and epitaxial growth of fully relaxed low-threading-dislocation-density SiGe buffers on Si substrates. He joined the Thomas J. Watson Research Center in 1991 to work on an advanced bi-CMOS project. He is currently with the Exploratory Devices and Processes group working on deep-submicron CMOS technology. Dr. Mii's current technical interests include 0.1-μm CMOS devices and process, SOI devices, and low-temperature CMOS.

David J. Frank *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (DFRANK at YKTVMV).* Dr. Frank received the B.S. degree from the California Institute of Technology, Pasadena, in 1977 and the Ph.D. degree in physics from Harvard University in 1983. Since graduation he has worked at the Thomas J. Watson Research Center, first as a postdoctoral fellow studying nonequilibrium superconductivity, and then as a research staff member modeling and measuring III–V devices. More recently he has been involved in exploring the limits of scaling through the modeling of innovative Si devices and in investigating the usefulness of energy-recovering CMOS logic and reversible computing concepts. His interests include superconductor and semiconductor device physics, modeling and measurement, circuit design, and percolation in two-dimensional systems. Dr. Frank is a member of the IEEE.

H.-S. Philip Wong *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (PWONG at YKTVMV).* Dr. Wong is a research staff member at the Thomas J. Watson Research Center. He received the B.Sc. (Hon.) degree in electrical engineering from the University of Hong Kong in 1982, the M.S. degree in electrical engineering from the State University of New York at Stony Brook in 1983, and the Ph.D. degree in electrical engineering from Lehigh University, Bethlehem, Pennsylvania, in 1988. He joined the Research Center that same year to work on various aspects of charge-coupled devices (CCDs) aimed at high-quality image capture for multimedia applications. In 1993, he joined the Silicon Technology Department to work on exploratory devices and processes for sub-0.1-μm CMOS. His present research interests are the physics and technology of the double-gate MOSFET, discrete random dopant effects in MOSFETs, hot-electron-induced photon emission, charge-based analog CMOS neural network hardware for image processing, and large-area crystalline silicon CMOS for projection displays. Dr. Wong's experience also includes electron device physics, device modeling, microelectronics process technology, and solid state imagers.

Douglas A. Buchanan *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York, 10598 (buchan@watson.ibm.com).* Dr. Buchanan received his B.Sc. and M.Sc. degrees in electrical engineering from the University of Manitoba, Winnipeg, Canada, in 1981 and 1982, respectively. In 1986 he received his Ph.D. from Durham University, Durham, England, with a dissertation entitled "Electronic Conduction in Silicon-Rich Thin Films." Upon receiving his Ph.D., Dr. Buchanan held a two-year postdoctoral fellowship in the Insulator Physics group at the

Thomas J. Watson Research Center, after which he spent three years in a CVD thin-film technology group in the IBM Microelectronics Division, East Fishkill, New York. Currently Dr. Buchanan works in the Exploratory Devices and Processes group at the Thomas J. Watson Research Center, on issues pertaining to the growth and characterization of thin dielectrics for sub-0.1-μm CMOS technology. He is a member of the American Physical Society and the Institute of Electrical and Electronics Engineers.

Shalom J. Wind *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (WIND at WATSON).* Dr. Wind is a research staff member in the Silicon Technology Department at the Thomas J. Watson Research Center. In 1987 he received a Ph.D. degree in physics from Yale University, where he studied electron quantum transport in nanostructures. He began his work at the Research Center in the Electron Beam Technology area in 1987. Since that time, Dr. Wind has been involved in high-resolution electron-beam lithography and processes, the fabrication of nanostructures and exploratory devices, and the study of ultrasmall devices at the limits of fabrication feasibility.

Stephen A. Rishton *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York, 10598 (RISHTON at YKTVMV).* Dr. Rishton received his Ph.D. in electronic engineering from the University of Glasgow, Scotland, in 1984, and his B.Sc. (Eng.) degree from University College, London, in 1980, also in electronic engineering. He has worked on the fabrication of various ultrasmall structures and electron-beam lithography since joining IBM at the Thomas J. Watson Research Center in 1984.

George A. Sai-Halasz *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (SAI at YKTVMV).* Dr. Sai-Halasz graduated in physics from Roland Eotvos Science University, Budapest, Hungary, in 1966. He received the Ph.D. degree in physics from Case Western Reserve University in 1972. During the following two years he held a postdoctoral fellowship at the Physics Department of the University of Pennsylvania. He joined the Thomas J. Watson Research Center in 1974. Dr. Sai-Halasz has made significant contributions in the areas of quantum solids, nonequilibrium superconductivity, and the physics and device aspects of semiconductor superlattices; he was one of the originators of the Type II superlattice system, and invented and developed a statistical modeling scheme for predicting radiation-induced soft-error rates in VLSI circuits. He has also worked in the area of scaling devices and circuits, demonstrating the possibility of a 0.1-μm FET technology. His most recent work is in the area of high-end processor design.

Edward J. Nowak *IBM Microelectronics Division, Burlington facility, Essex Junction, Vermont 05452 (ejnowak @ vnet.ibm.com).* Dr. Nowak is a senior engineer in the Logic Device Design Department in Essex Junction, Vermont. Since joining IBM there in 1981, he has worked on advanced memory and logic device integration and design projects. Dr. Nowak received a B.S. in physics from M.I.T. in 1973 and M.S. and Ph.D. degrees in physics from the University of Maryland in 1976 and 1979, respectively.