Impact of Self-Heating Effect on Long-Term Reliability and Performance Degradation in CMOS Circuits

Oleg Semenov, *Member, IEEE*, Arman Vassighi, *Associate Member, IEEE*, and Manoj Sachdev, *Senior Member, IEEE*

***Abstract*—As the technology feature size is reduced, the thermal management of high-performance very large scale integrations (VLSIs) becomes an important design issue. The self-heating effect and nonuniform power distribution in VLSIs lead to performance and long-term reliability degradation. In this paper, we analyze the self-heating effect in high-performance sub-0.18-μm bulk and silicon-on-insulator (SOI) CMOS circuits using fast transient quasi-dc thermal simulations. The impact of the self-heating effect and technology scaling on the metallization lifetime and the gate oxide time-to-breakdown (TBD) reduction are also investigated. Based on simulation results, an optimized clock-driver design is proposed. The proposed layout reduces the hot-spot temperature by 15 °C and by 7 °C in 0.09-μm SOI and bulk CMOS technologies, respectively.**

***Index Terms*—CMOS technology scaling, long-term reliability, performance degradation, self-heating effect.**

I. Introduction

TRANSISTOR scaling is the primary factor in achieving high-performance microprocessors and memories. A 30% reduction in CMOS integrated-circuit (IC) technology node scaling has: 1) reduced the gate delay by 30%, allowing an increase in the maximum clock frequency of 43%; 2) doubled the device density; 3) reduced the parasitic capacitance by 30%; and 4) reduced the energy and active power per transition by 65% and 50%, respectively [1]–[3].

The overall power consumption of microprocessors can be categorized into two types: active and passive. The logic gate switching (active) power density ($P_{\mathrm{sw}} \sim C_{\mathrm{gate}} V_{\mathrm{DD}}^2/\tau$, where C_{gate} is the gate capacitance, V_{DD} is the power supply, and τ is the logic gate delay) should remain constant with classic CMOS scaling. However, P_{sw} obtained from the industrial data increases with the technology scaling because of the lag in V_{DD} reduction, which is only partially mitigated by a reduction in C_{gate} [4]. In other words, if a die size is kept constant, to add more function with scaling, the overall switching power must increase. The standby leakage current density increases exponentially as the transistor channel length is decreased. This follows from the demand that V_{TH} (threshold voltage of MOSFET) decreases with V_{DD} scaling and the fact that the leakage current increases exponentially with a V_{TH} reduction. Furthermore, in the most recent CMOS generations, tunneling current through the gate oxide has become a significant part of the overall leakage current. This current causes the additional passive power consumption in CMOS circuits. Note that passive power, unlike active power, is dissipated by all CMOS circuits all the time, whether or not they are actively switching. The total power consumption trend of microprocessors with technology scaling is presented in Fig. 1 [2]. It can be seen that the power consumption is increased by $\sim 2.3\times$ with scaling from the 0.18-μm CMOS technology to the 0.09-μm CMOS generation.

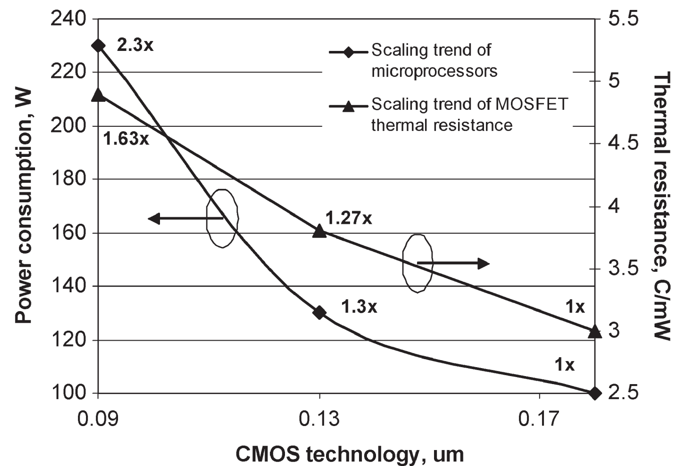

Fig. 1. Power consumption of microprocessors [2] and MOSFETs thermal resistance [6] versus CMOS technology scaling.

Recently, a relationship between the thermal resistance (R_{TH}) of an MOSFET and its geometrical parameters was derived using a three-dimensional heat flow equation [5]. It was shown (Fig. 1) that R_{TH} is increased by $\sim 1.63\times$ with scaling from the 0.18-μm CMOS to the 0.09-μm CMOS [6]. Due to the increase in power consumption and MOSFET thermal resistance, the average junction temperature (T_j) of high-performance microprocessors is also increased with technology scaling [7]. Currently, high-performance microprocessors have average power densities ranging from 10 to 100 W/cm^2 or even more. Hence, the average junction temperature often exceeds 100 °C [8]. This problem is aggravated by the fact that the microprocessor power distribution is far from uniform, with small regions of the CPU dissipating a significant fraction of the

Manuscript received June 2, 2005; revised December 8, 2005.

O. Semenov and M. Sachdev are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: osemenov@vlsi.uwaterloo.ca).

A. Vassighi was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada. He is now with Intel Corporation, Hillsboro, OR 97124 USA.

Digital Object Identifier 10.1109/TDMR.2006.870340

1530-4388/$20.00 © 2006 IEEE

total power. Nonuniform power dissipation leads to a nonuniform temperature distribution on the die surface with localized high-temperature regions (hot spots). These hot spots affect the product long-term reliability and processor performance and yield.

In this paper, we analyze the impact of CMOS technology scaling on the temperature increase of CPU hot-spot regions, where high-performance circuits are located. Using long-term reliability models, we estimate the degradation of time-to-failure (TTF) of metallization and the time-to-breakdown (TBD) of gate oxides in sub-0.18-μm high-performance circuits. The effect of the elevated temperature on performance degradation is also investigated.

This paper is organized as follows. In Section II, we review the thermal-management aspects of high-performance circuits. The self-heating effect in bulk and silicon-on-insulator (SOI) CMOS technologies, applied for the advanced microprocessors fabrication, is analyzed in Section III using fast transient quasi-dc thermal simulations. The design optimization of high-performance CMOS drivers for heating-effect reduction is described in Section IV. The impact of technology scaling on the self-heating effect and the performance-degradation and long-term-reliability projections are discussed in Sections V and VI, respectively. Finally, the conclusions are summarized in Section VII.

II. Thermal-Management Background

The primary function of the thermal-management system is to monitor the temperature of the system and regulate the system's operating temperature. A key aspect of thermal management is ensuring that the junction temperature of the microprocessor does not exceed the operating specification. A small increase of 10 °C–15 °C in the junction temperature may result in $\sim 2\times$ reductions in the lifespan of the device [9]. While T_a represents the ambient temperature for an IC, the relationship between the ambient temperature and the average junction temperature (T_j) for a very large scale integration (VLSI) is often described as [10]

$$T_j = T_a + P \times R_{\mathrm{TH}}. \quad (1)$$

In (1), P is the total power consumption of the chip and R_{TH} is the effective junction-to-ambient thermal resistance. In a general case, the R_{TH} includes the junction-to-ambient thermal resistance of the chip, the package, and the heat sink. Since the overall power consumption of microprocessors and the MOSFET thermal resistance are increased with scaling, package and heat-sink thermal resistances should be reduced with technology scaling to keep the average and hot-spot junction temperatures at the acceptable level. It was predicted that a reduction of approximately 22% in package thermal resistance per technology generation is required to just to compensate the increased power density with technology scaling [11]. The intensive industrial investigations have shown that the heat-sink thermal resistance can be scaled from 0.3 °C/W for current microprocessors to 0.13 °C/W for future chips [12]. Note that the control of the hot-spot junction temperature is more critical in comparison with the control of the average junction temperature, since the hot-spot power density reaches 250 W/cm^2 for 3-GHz microprocessors [11].

III. Self-Heating Effect in High-Performance Circuits

Historically, high-performance circuits such as microprocessors have been fabricated using the bulk CMOS technology, since this technology is cheaper than the SOI technology. However, it has been recently shown that a partially depleted (PD) SOI CMOS technology can provide significant performance advantages over bulk CMOS in sub-0.18-μm generations [13], [14]. Since bulk and SOI CMOS technologies are currently used for high-performance circuits, both are considered in this paper.

A. *SOI and Bulk Multifinger MOSFETs*

Typically, large MOSFETs are designed in multifinger configurations. In our research, we used the two-dimensional "Sequoia Device Designer" simulation software [15] for MOSFET physical-structure generation and thermal simulations. This simulator has built-in device synthesis, mesh generation, device simulation, circuit-device mixed-mode simulation, and lattice self-heating simulation modules. The cross sections of the generated bulk and SOI multifinger devices are shown in Fig. 2. We can study the thermal coupling effect between fingers in multifinger MOSFETs by varying the separation between the fingers (i.e., fingers 2 and 3 in Fig. 2).

The model of the bulk multifinger transistor generated by the Sequoia software was calibrated using the process and device data from TSMC (Taiwan Semiconductor Manufacturing Company) 0.18-μm-CMOS-technology file. For the SOI multifinger transistor model, the bulk silicon substrate was substituted with the SOI substrate. To estimate the impact of silicon-film thickness (D_{Si}) and buried-oxide (BOX) thickness (D_{BOX}) on the self-heating effect in high-performance circuits, the following cases were considered:

1) bulk MOSFET;
2) SOI MOSFET with $D_{\mathrm{Si}} = 0.5\ \mu\mathrm{m}$ and $D_{\mathrm{BOX}} = 0.25\ \mu\mathrm{m}$;
3) SOI MOSFET with $D_{\mathrm{Si}} = 0.5\ \mu\mathrm{m}$ and $D_{\mathrm{BOX}} = 1.5\ \mu\mathrm{m}$;
4) SOI MOSFET with $D_{\mathrm{Si}} = 0.03\ \mu\mathrm{m}$ and $D_{\mathrm{BOX}} = 1.5\ \mu\mathrm{m}$.

To emulate the impact of package and heat-sink thermal resistances on the self-heating effect in high-performance circuits, an effective thermal resistance of 0.3 °C/W [11] was connected in series with the silicon substrate of the device structures shown in Fig. 2.

B. *Clock-Driver Thermal Simulations*

The power dissipation due to the clock distribution network is dominant in high-performance microprocessors. For example, the main clock drivers consume 40% of the total chip power in the Alpha 21064 and Alpha 21164 microprocessors [16]. In the POWER4 microprocessor, approximately 70% of the

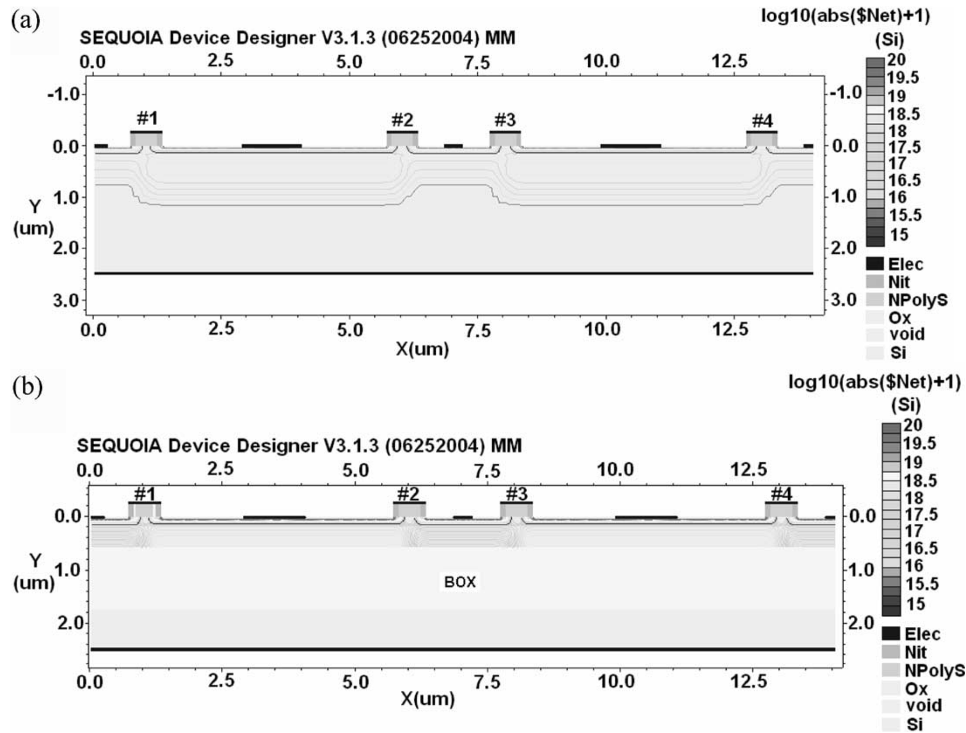

Fig. 2. Cross section of a four-finger 0.18-μm MOSFET in (a) bulk and (b) SOI technologies.

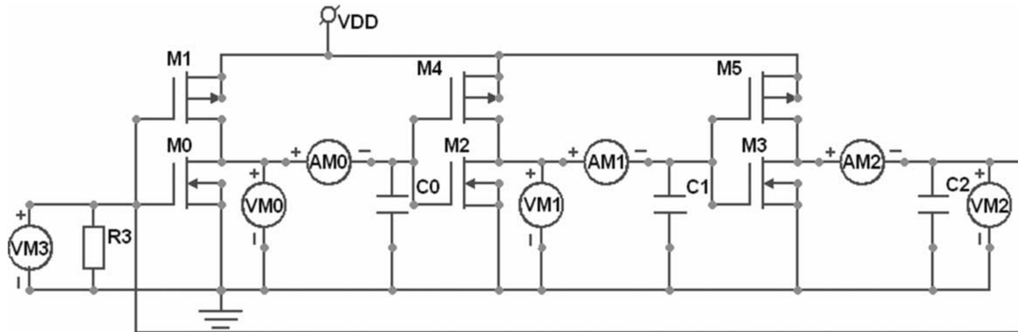

Fig. 3. Schematic of simulated three-stage ring oscillator.

overall power is consumed in the clock distribution network and latches [17]. The power density of clock drivers is extremely high as they are required to generate and distribute a "high-quality" signal with low clock skew through the chip. As such, the thermal management of clock drivers is a major concern.

1) Fast Transient Thermal Simulations: To emulate the self-heating effect in a high-performance CMOS clock-driver network, we simulate a three-stage ring oscillator that consumes approximately the same power as a typical clock driver implemented in 0.18-μm CMOS [18]. The typical overall power consumption of a clock driver in this technology is $\sim$ 48.6 mW, assuming a 15-pF clock load, a 1.8-V supply voltage, and 1-GHz operational frequency [18]. The schematic of the simulated three-stage ring oscillator is given in Fig. 3.

For electrical and thermal simulations of the ring oscillator, we use the mixed-mode (device-circuit) module of Sequoia. Each transistor in the ring oscillator was implemented as a physical structure shown in Fig. 2. For the bulk CMOS technology, the total width of the NMOS transistors (M0, M2, and M3) was 150 μm, and the width of PMOS transistors (M1, M4, and M5) was 300 μm. The capacitance loads (C0, C1, and C2) were 10 pF each. The simulated operating frequency of the ring oscillator found from simulations was approximately 3.3 GHz, and the average power consumption (P_{av}) was approximately 70 mW $(I_{\mathrm{av}} \sim 40$ mA). In the case of the SOI technology (case 2 mentioned above), the transistor sizes were optimized to provide the same power consumption as that obtained in the bulk CMOS technology. This allows for an effective analysis of the increasing of self-heating effect between the bulk CMOS and the SOI. The results of the fast transient thermal simulations are shown in Fig. 4. In this figure, the peak temperature in hot-spot regions of the analyzed ring oscillator is shown as a function of the simulating time. The hot-spot regions of the analyzed circuit are located in the drain junctions of NMOS transistors as shown in the inset of Fig. 4.

From Fig. 4 ,we can conclude that high-performance circuits based on the SOI technology have a significantly higher self-heating effect. The second conclusion is that the thermal time

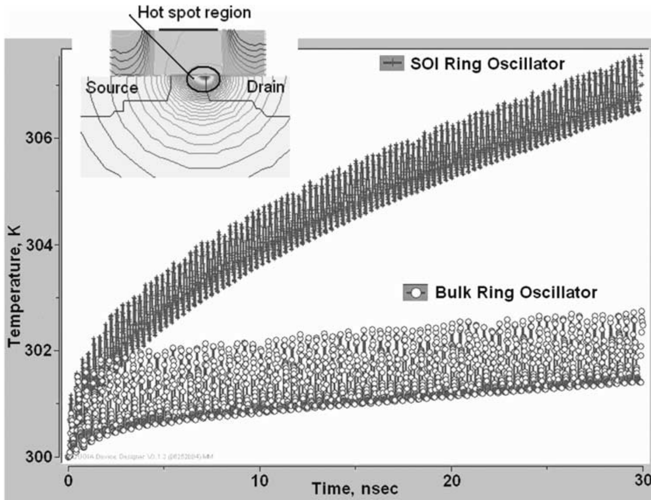

Fig. 4. Fast transient thermal-simulation results of bulk and SOI three-stage ring oscillators.

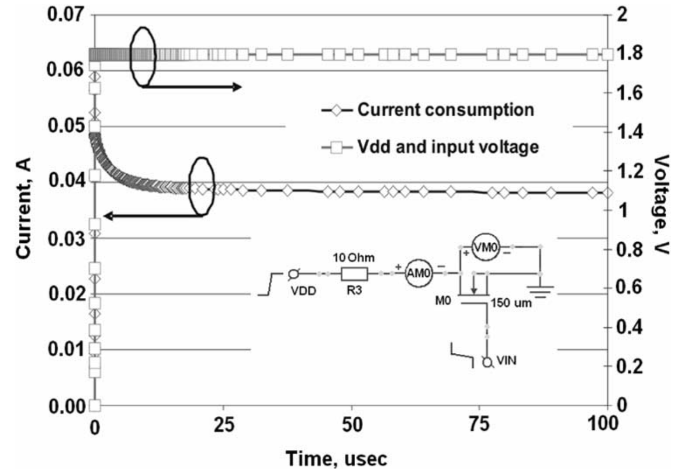

Fig. 5. Voltage waveform used for quasi-dc thermal simulations and current consumption waveform of resistor–transistor inverter.

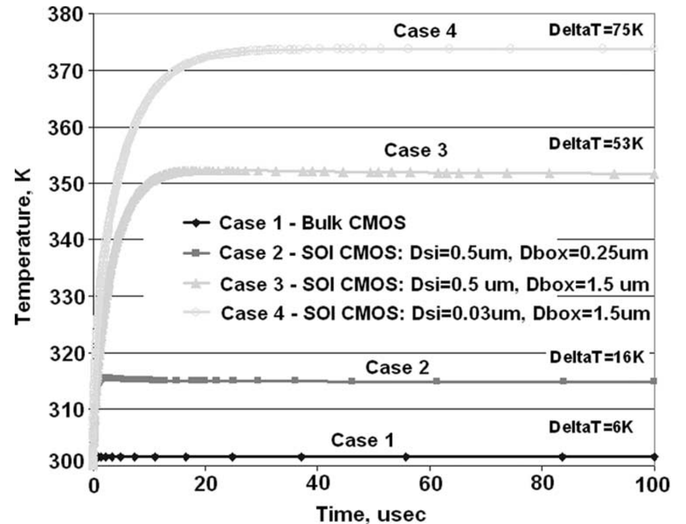

Fig. 6. Quasi-dc-thermal-simulation results of clock drivers: Peak temperature versus simulation time.

constant $(\tau_{\rm th})$ is much longer than the 30 ns of the ring-oscillator operating time. The obtained curves are not saturated during this time. Different blocks in the microprocessor chip have the thermal time constants from the order of tens to hundreds of microseconds [19]. It should be noted that the 30 ns of fast transient thermal simulations took approximately 11 h of physical simulation time on a 3.2-GHz Pentium-4 microprocessor.

2) Quasi-dc Thermal Simulations: Since the fast thermal simulations cannot be used for hot-spot-temperature detection in a reasonable simulation time, we performed quasi-dc thermal simulations. The peak junction temperature increase over the ambient temperature in a transistor is directly proportional to the total power dissipation of an MOSFET and its thermal impedance. Here, we should note that the transient thermal impedance $(Z_{\rm th})$ and the steady-state thermal resistance $(R_{\rm TH})$ of an MOSFET are not equal in general. However, if the actual operating time is significantly longer than the thermal time constant $(\tau_{\rm th})$, then one may assume that the chip is under a thermal steady state, i.e., $R_{\rm TH} = Z_{\rm th}$ (time $\gg \tau_{\rm th}$) [20] and the quasi steady-state thermal analysis can be used. Recently performed fast transient thermal analysis of high-performance circuits have shown that different blocks of microprocessors can have the thermal time constants in the range of tens to hundreds microseconds under normal operating conditions [19]. In reality, a high-performance chip is operating for much longer than tens or hundreds of microseconds; therefore, the approximation of a quasi-dc thermal case can be used. Note that the equivalent circuit used for a quasi-dc thermal analysis should have the same power consumption at quasi-dc operating conditions as that of an original circuit at normal operating conditions.

To perform the quasi-dc thermal simulations, we applied a step signal with a 50-ps rise time to $V_{\rm DD}$ and $V_{\rm IN}$ terminals of a resistor–transistor inverter as shown in the inset of Fig. 5. In order to dissipate approximately 40 mA of dc current, the resistor and the transistor sizes were chosen to be 10 Ω and 150 μm/0.18 μm, respectively. This circuit consumes approximately the same power as one stage of the ring oscillator described in Section III-B1. Fig. 5 illustrates the input voltage and the current through the inverter.

The quasi-dc-thermal-simulation results of the equivalent inverter for both bulk CMOS and SOI technologies are presented in Fig. 6. From this figure, we can conclude that the peak temperature increase in the hot-spot regions of the clock driver implemented in a bulk CMOS technology (case 1) is ~ 6 °C over the room temperature. Clock drivers implemented in the SOI technology have significantly higher self-heating effect: ~ 16 °C for case 2; ~ 53 °C for case 3; and $\sim$ 75 °C for case 4.

Increasing the BOX thickness by $6\times$ (cases 3 and 4) and reducing the silicon-film thickness by $\sim 17\times$ (case 4) result in a stronger self-heating effect of $3.3\times$ and $1.4\times$, respectively, compared to case 2. In practice, thick BOX SOI technologies are used for smart-power applications. Note that ultra-thin silicon films (case 4) are typically used in fully depleted (FD) SOI technologies. Hence, FD SOI technologies have a stronger

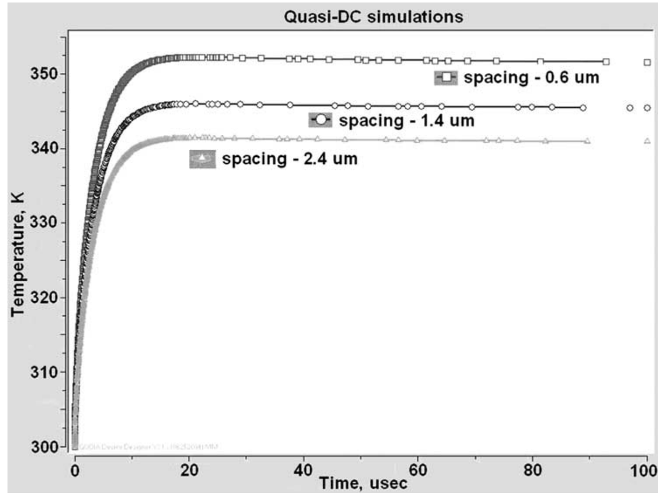

Fig. 7. Peak temperature in the clock driver implemented in the SOI technology (case 3) versus the interfinger spacing.

TABLE I
SELF-HEATING EFFECT VERSUS INTERFINGER SPACING FOR DIFFERENT BULK AND SOI TECHNOLOGIES

Bulk & SOI Tech.	X_{sp}=0.6 um	X_{sp}=1.4 um	X_{sp}=2.4 um	$\Delta T_{Xsp=0.6}/\Delta T_{Xsp=2.4}$
Case 1	ΔT=6°C	ΔT=4.5°C	ΔT=4°C	1.5x
Case 2	ΔT=16°C	ΔT=13.5°C	ΔT=12°C	1.3x
Case 3	ΔT=53°C	ΔT=45°C	ΔT=40°C	1.3x
Case 4	ΔT=75°C	ΔT=65°C	ΔT=55°C	1.4x

self-heating effect than PD ones and demand careful thermal management.

IV. DESIGN OPTIMIZATION OF CMOS CLOCK DRIVER

Since the channel width of transistors in high-performance clock drivers is significantly large, these devices are often laid out in a multifinger fashion. To accurately study the impact of the layout on the thermal coupling effect in these devices, we perform quasi-dc thermal simulations with different spacings (X_{sp}) between transistor fingers. All simulation results presented in Section III were obtained with the minimum design rule spacing (0.6 μm for the 0.18-μm CMOS technology) between fingers 2 and 3 (see the Fig. 2).

For the analysis of the thermal coupling effect, we perform quasi-dc thermal simulations on equivalent inverters with different spacings between fingers 2 and 3 in a four-finger MOSFET. All cases described in Section III-A are simulated with varying finger spacing. Fig. 7 illustrates thermal-simulation results for case 3 (peak temperature versus simulation time at a different interfinger spacing). The complete simulation results for bulk and other SOI technologies are summarized in Table I. From this table, we can conclude that the increase of the interfinger spacing from 0.6 to 2.4 μm can reduce the self-heating effect by $1.3\times$ to $1.5\times$.

The thermal coupling effect is illustrated in Fig. 8. In Fig. 8(a), we can see that the thermal-distribution profiles have a stronger overlap when fingers have a minimum spacing; while in Fig. 8(b) with a spacing of 2.4 μm, the coupling effect is much weaker. As a result, the peak temperature with the minimum spacing is higher than the peak temperature with a 2.4-μm spacing.

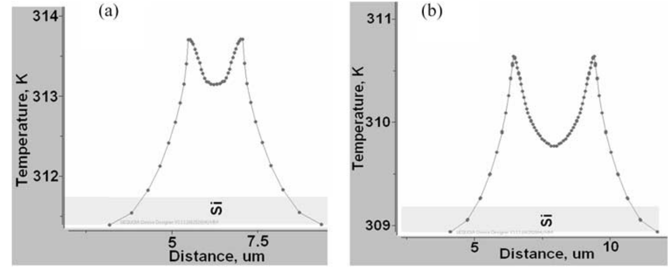

Fig. 8. Illustration of the thermal coupling effect (case 2). (a) Spacing between fingers is 0.6 μm. (b) Spacing between fingers is 2.4 μm.

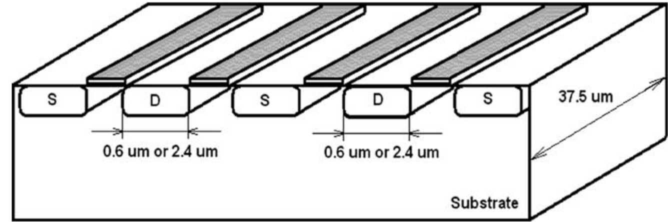

Fig. 9. Drain area and capacitance of a four-finger MOSFET.

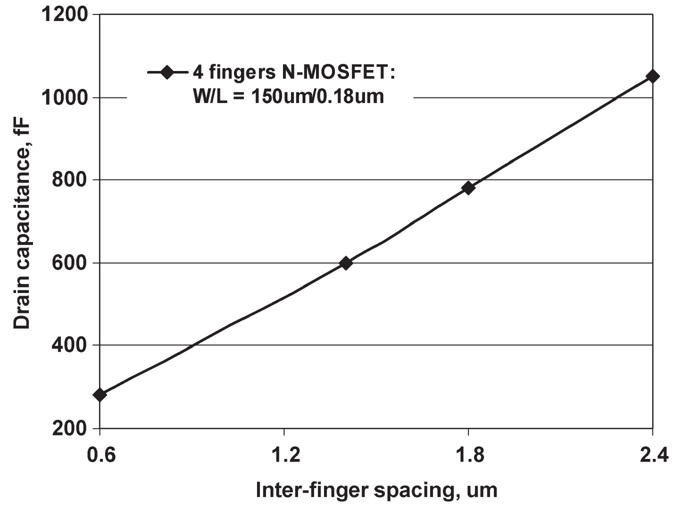

Fig. 10. Drain capacitance versus interfinger spacing.

A. Interfinger Spacing Optimization

From Fig. 8, we can conclude that the increased interfinger spacing in transistors may reduce the self-heating effect in CMOS clock drivers. However, the increase in the drain area causes an increase in the drain capacitance in bulk and PD SOI technologies, where drain and source regions do not touch the BOX. This is shown in Figs. 9 and 10. The dependence of the drain capacitance versus the drain area was extracted from simulations in Cadence. As the finger spacing increases, the capacitance also increases; thus, the performance of the CMOS clock driver degrades. Consequently, it is necessary

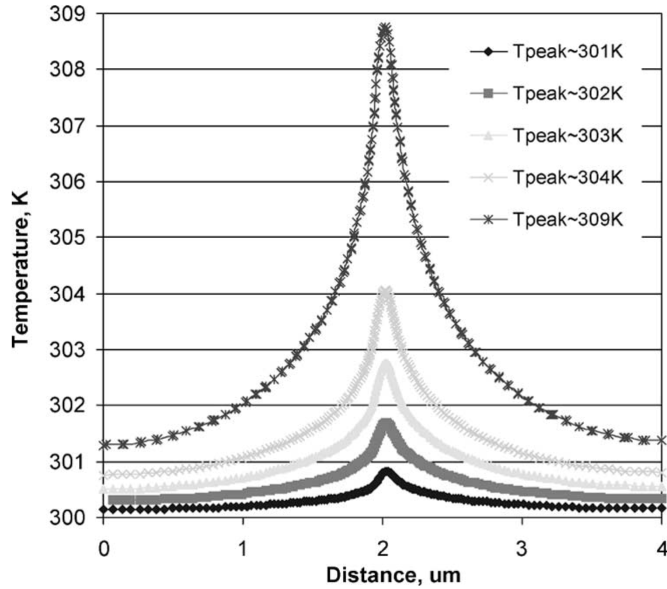

Fig. 11. Temperature distributions in a four-finger N-MOSFET due to the self-heating effect (case 1).

to find the optimal value of the interfinger spacing to reduce the self-heating effect and to avoid the performance degradation. The typical capacitance load of CMOS clock drivers in 0.18-μm CMOS technology is approximately 10–15 pF [18]. Hence, to avoid the significant performance degradation of the clock driver, the total drain capacitance should be at least $10\times$ less than the capacitance load, since these capacitances are connected in parallel.

Fig. 10 shows that the limit of the interfinger spacing is 2.4 μm ($4\times$ the minimum design rule value for 0.18-μm CMOS technology). In this case, the total drain capacitance is less than the load capacitance of the clock driver by $10\times - 15\times$. The second issue is the reduction of the thermal coupling effect in the multifinger device if the 2.4-μm interfinger spacing is used. The temperature distributions for a four-finger NMOS (case 1) obtained from the quasi-dc thermal simulations are shown in Fig. 11. A curve in Fig. 11 depicts the temperature distribution for a single finger with respect to the distance from the heat source. Different curves show the thermal distribution as a function of the power dissipation. Using the curve-fitting tool in Origin [21], we estimate that the temperature distribution in the transistor can be best described by the pseudoVoigt distribution. The Gaussian distribution provides the worse curve fitting as compared to the pseudoVoigt distribution [see Fig. 12(a)]. From the curve-fitting analysis, the widths of temperature distributions (Z) were extracted as shown in Fig. 12(b). From this graph, it was found that the ($3 \times Z$) value corresponds to $\sim 80\%$ of the temperature reduction from the maximum value. The curve-fitting analysis was performed for all temperature profiles shown in Fig. 11. The same analysis was done for the PD SOI four-finger NMOS transistor (case 2). The extracted ($3 \times Z$) values for bulk and SOI technologies as a function of peak temperatures are presented in Fig. 13. Equations (2) and (3) can be used to analytically describe the obtained dependences for bulk and SOI technologies, respectively.

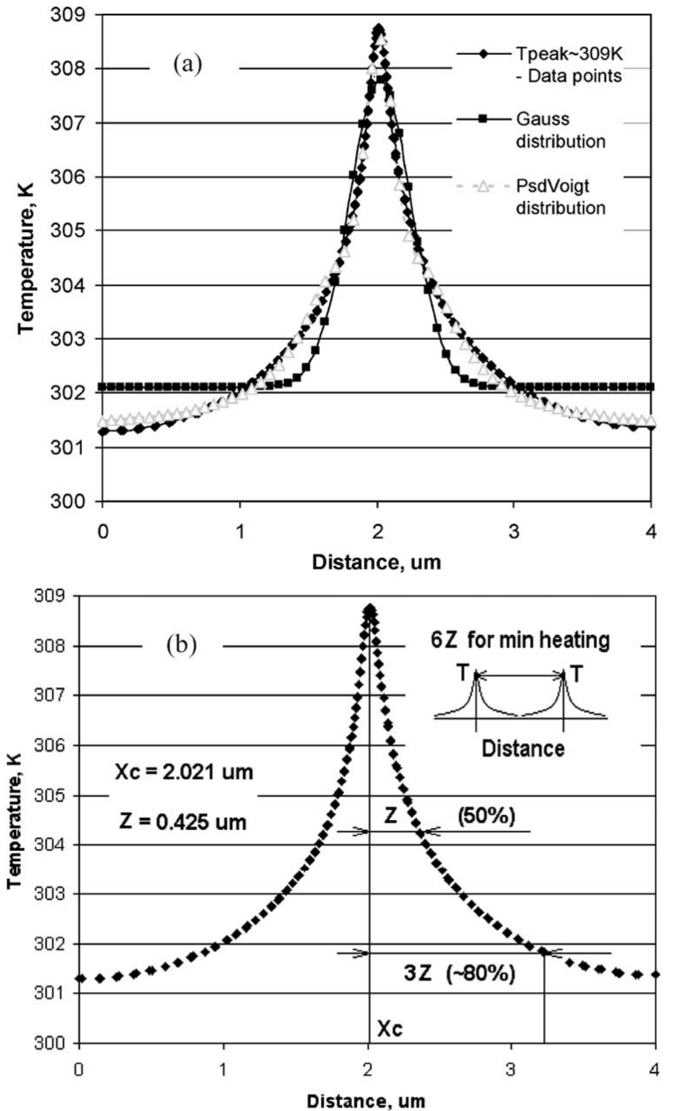

Fig. 12. (a) Results of the curve-fitting analysis and (b) parameters of the pseudoVoigt distribution extracted from the Origin software.

They were obtained using the curve-fitting analysis performed in Origin.

$$\begin{aligned} Y &= (a + b \times x)^{-\frac{1}{c}} \\ a &= -\ 323.994 \\ b &= 1.081 \\ c &= -\ 9.357 \end{aligned} \tag{2}$$

$$\begin{aligned} Y &= a|x - b|^{c} \\ a &= 0.589 \\ b &= 300.478 \\ c &= 0.103. \end{aligned} \tag{3}$$

In these equations, x is the absolute peak temperature and Y is the $(3Z)$ value of the pseudoVoigt distribution. Using (2) and (3), we can calculate that the situation where the hot spot dissipates 80% of the thermal energy required for the runaway condition corresponds to a finger spacing of ~ 2.2 and ~ 1.3 μm for bulk and SOI technologies, respectively. It is assumed that a catastrophic failure is the melting of the metal

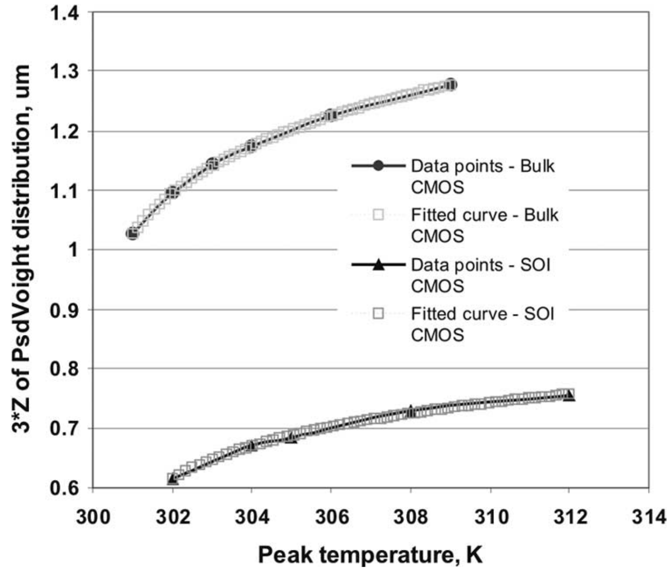

Fig. 13. $3Z$ values of pseudoVoigt distributions versus the peak temperature for bulk (case 1) and SOI (case2) technologies.

contact of the gate electrode, which happens if the hot spot has the peak temperature close to the temperature of silicon melting ($\sim$ 1683 K). Hence, to increase the robustness of multifinger devices, the interfinger spacing $(6Z)$ should be 4.4 μm for the bulk CMOS technology and 2.6 μm for the SOI CMOS technology. Previously, using the Gauss distribution for temperature profiles, it was found that the interfinger spacing should be $\sim$ 5 μm for a reliable electrostatic-discharge (ESD) design in bulk CMOS technologies [22]. In the case of SOI CMOS technologies, the optimal interfinger spacing, which is necessary to avoid the thermal coupling effect should be $\approx 5\sqrt{K_{\rm Si} D_{\rm Si} D_{\rm BOX}}/K_{\rm BOX}$ [23], where $K_{\rm Si}$ and $K_{\rm BOX}$ are the thermal conductivities of silicon and buried silicon dioxide, respectively, and $D_{\rm Si}$ and $D_{\rm BOX}$ are the thicknesses of silicon and buried silicon dioxide films, respectively. The thermal conductivity of high-purity monocrystalline bulk silicon has a temperature dependence $K_{\rm Si} = K_0 \times (300/T)^{4/3}$, where $K_0 = 154.86 \text{ W} \times K - 1 \times \text{m}^{-1}$ is the thermal conductivity of silicon at room temperature [24]. At room temperature, the thermal conductivity of the n-doped silicon is $K_0 = 29 \text{ W} \times \text{K}^{-1} \times \text{m}^{-1}$ [25]. The thermal conductivity of silicon dioxide with a thickness of 0.2 μm is approximately $1.2 \text{ W} \times \text{K}^{-1} \times \text{m}^{-1}$ at 1423 K [26]. Hence, at the failing temperature (the temperature of silicon melting, $\sim$ 1683 K), the calculated optimal interfinger spacing for the SOI technology (case 2) is approximately 2.8 μm. Since our results for the optimal interfinger spacing for bulk and SOI technologies (4.4 and 2.6 μm, respectively) are close to the previously published and analytically calculated ones (5 μm for bulk CMOS and 2.8 μm for SOI CMOS), we can conclude that our approach to the analysis of thermal simulations is correct.

To minimize the thermal coupling effect in multifinger transistors, the interfinger spacing should be at least $6Z$. The peak junction temperature in hot-spot regions should not exceed 150 °C in the worst case scenario due to the long-term reliability requirements (see Section V). Using (3) and (4), the $6Z$ values estimated for this temperature are 3.4 and 1.9 μm for bulk and SOI CMOS technologies, respectively. Note that the 4$\times$ minimum design rule value for the 0.18-μm CMOS technology, which is optimal for the interfinger spacing in high-performance clock drivers, is 2.4 μm. Hence, the thermal coupling effect cannot be avoided completely. However, the interfinger spacing of 4$\times$, where $\times$ is the minimum design value for the given CMOS technology, provides the reasonable compromise between the performance degradation of the high-speed clock driver and the self-heating-effect reduction. Note that the 4$\times$ minimum design rule value for the 0.09-μm CMOS technology is only 1.2 μm. Furthermore, the power density increases with scaling [4]; therefore, one may expect a higher degree of thermal coupling with scaling of technology.

V. Technology Scaling and Self-Heating Effect in High-Performance Circuits

In this section, we consider the impact of technology scaling on the self-heating effect (temperature increase in the hot-spot region) in VLSIs. We assume that the power consumption of high-performance clock drivers has the same trend with scaling as a total power consumption of microprocessors, which is shown in Fig. 1. Then, the self-heating-effect scaling from the 0.18-μm CMOS technology to the 0.13-μm CMOS technology can be calculated by the following equation:

$$\begin{aligned}\Delta T_{0.13} &= \Delta T_{0.18} \times \frac{P_{0.13}}{P_{0.18}} \times \frac{R_{\rm TH0.13}}{R_{\rm TH0.18}} \\ &= \Delta T_{0.18} \times P_n \times R_n. \end{aligned} \tag{4}$$

In (4), $\Delta T_{0.18}$ is the self-heating effect for the 0.18-μm CMOS technology, obtained from quasi-dc thermal simulations; R_n is the normalized device thermal resistance with scaling from the 0.18-μm technology to the 0.13-μm technology; and P_n is the normalized power consumption with scaling from the 0.18-μm technology to the 0.13-μm technology. R_n and P_n can be found from Fig. 1. The same approach can be used for calculation of the self-heating-effect scaling from a 0.18- to a 0.09-μm technology. In this paper, we considered two clock-driver designs with the minimum design spacing (0.6 μm) between fingers 2 and 3 (Fig. 2) and with the 4$\times$ larger interfinger spacing (2.4 μm). The calculation results are shown in Fig. 14. Note that the junction temperature in hot-spot regions should not exceed the long-term reliability limit. For the silicon technology this limit is 150 °C [27]. It is assumed that the gate oxide wear-out and the metallization degradation become unacceptable if the junction temperature exceeds the long-term reliability limit. The average junction temperature of microprocessors implemented in the 0.18-μm CMOS technology is approximately 90 °C [28]. Hence, the junction temperature increase in hot-spot regions should not exceed 60 °C from the long-term reliability point of view. If we assume that the average junction temperature of a CPU is not increased with scaling, due to the reduction of package and heat-sink thermal resistances and due to the usage of advanced cooling techniques, it leaves a temperature margin of 60 °C for all considered technologies. From Fig. 14(a), we can conclude that the peak temperature increase in SOI high-performance circuits

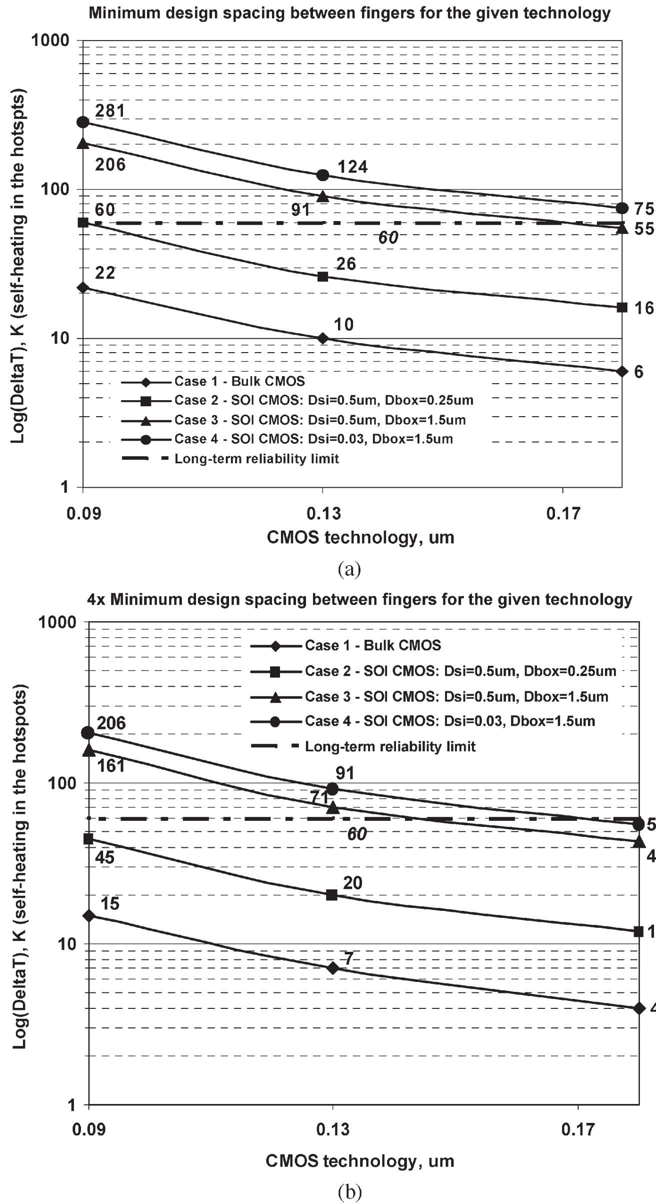

Fig. 14. Self-heating effect versus technology scaling. (a) Minimum design spacing and (b) 4× minimum design spacing.

(case 2) implemented in the 0.09-μm CMOS technology becomes critical if the minimum design interfinger spacing is used. To satisfy long-term reliability requirements, higher (4×) interfinger spacing should be applied [see Fig. 14(b)]. Note that SOI technologies with thick BOX layers (cases 3 and 4), which are typically used for smart-power applications, cannot be used for high-performance circuit fabrication due to the strong self-heating effect.

A. *Performance Degradation Due to the Self-Heating Effect*

The proper thermal design is important not only for reliability aspects, but also for the circuit performance optimization. The increase of the junction temperature in hot-spot regions of high-performance circuits may have a significant impact on performance degradation (i.e., operating frequency). To study this effect, we simulate the three-stage ring oscillator (Fig. 3) in

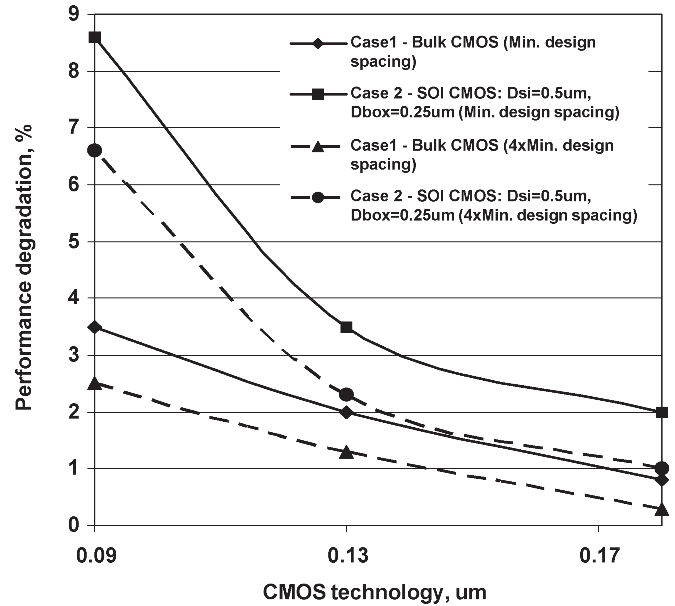

Fig. 15. Performance degradation of a three-stage ring oscillator due to the self-heating effect versus technology scaling.

Cadence using HSPICE simulator and determined the operating frequency at a different interfinger spacing. We performed these simulations at the ambient temperature equal to the room temperature and the peak junction temperature shown in Fig. 14. The performance degradation was defined as a relative reduction of the operating frequency due to the self-heating effect in comparison with the operating frequency at room temperature. It was assumed that the parasitic drain capacitance in multifinger transistors with 4×-increased interfinger spacing is negligibly small in comparison with the total capacitance load of the clock driver. This is a reasonable assumption because, in practice, a sector clock driver in an H-tree clock distribution system has a capacitance load of approximately 150 pF, and this value typically increases with CMOS technology scaling [29], [30]. The mentioned microprocessor clock distribution system implemented in 0.25-μm CMOS technology has an operating frequency equal to 400 MHz. Since SOI technologies with thick BOX layers (cases 3 and 4) are not used for high-performance circuit fabrications, we consider only bulk (case 1) and thin BOX SOI (case 2) technologies. The obtained simulation results are presented in Fig. 15.

From Fig. 15, we can conclude that starting from the 0.13-μm CMOS technology node, the performance degradation due to the self-heating effect becomes significant. Note that the speedup offered by the PD SOI technology over the bulk CMOS technology shows a declining trend with scaling because of the "history effect" in PD SOI transistors. It has been shown that an arithmetic logic unit (ALU) implemented in 0.13-μm PD SOI technology gets a speed up approximately 16% over bulk CMOS. The speed up of an ALU implemented in 0.09-μm PD SOI technology is estimated at approximately 12% [13]. In addition, in SOI, the self-heating-related performance degradation is expected to increase with scaling, which may dramatically diminish the performance advantage of the SOI technology over the bulk technology (Fig. 15). The second

conclusion from Fig. 15 is that the optimized clock-driver design with increased interfinger spacing can reduce the performance degradation due to the self-heating effect by 1.3×–1.4× in the 0.09-μm technology.

VI. Projections of Long-Term Reliability Degradation Due to Self-Heating Effect

Generally, device and interconnect failure mechanisms are classified into three main groups: 1) electronic stress failures (electrical-overstress (EOS) and ESD failures); 2) intrinsic failures (gate oxide breakdown, interconnect degradation, crystal defects, and ionic contamination); and 3) extrinsic failures (die-attachment and chip-packaging failures). In this paper, we consider the impact of the self-heating effect on two major types of intrinsic failures that are accelerated by self-heating.

A. Effect of Self-Heating on Metallization Lifetime

The interconnect electromigration lifetime under quasi-dc operating conditions can be estimated from the well-known Black's equation [31]

$$\mathrm{TTF} = \frac{A\exp\left(\frac{E_a}{kT}\right)}{J_{\mathrm{ave}}^2}. \tag{5}$$

In (5), TTF is the TTF of interconnect, A is the process-dependent constant, k is the Boltzmann's constant, E_a is the activation energy ($\sim$ 0.7–0.9 eV for Cu interconnections), T is the interconnect absolute temperature, and J_{ave} is the average current density. From (5), we can see that the TTF due to electromigration is related exponentially to the reciprocal temperature. Therefore, the increase in the metal-line temperature can cause a significant reduction in the TTF. Using this equation, we estimated the TTF reduction (TTF ratio) of metal layers in clock drivers with technology scaling due to the self-heating effect. The calculations were performed assuming that the interconnect temperature is equal to the peak junction temperature shown in Fig. 14. We considered only bulk and PD SOI CMOS technologies that are practically used for high-performance circuit fabrications. The obtained results are shown in Fig. 16. These results show that the increased interfinger spacing in MOSFETs can reduce the lifetime degradation of metal lines by 2.7× for the 0.9-μm PD SOI CMOS technology and by 1.8× for the 0.9-μm bulk CMOS technology. However, in spite of an increased spacing, the lifetime degradation of metallization for the PD SOI CMOS technology is still extremely high ($\sim 34\times$). Hence, for sub-0.09-μm PD SOI CMOS technologies, advanced cooling techniques are required to prevent the long-term reliability problems in interconnects due to the self-heating effect.

Note that the calculation results shown in Fig. 16 were obtained assuming that the average current density (J_{ave}) of the interconnects in (5) is independent of the temperature. The leakage power contribution is exponentially increased with scaling and also has strong temperature dependence (Fig. 17). Starting from the 0.09-μm CMOS technology, the leakage power becomes a significant contributor of the overall power

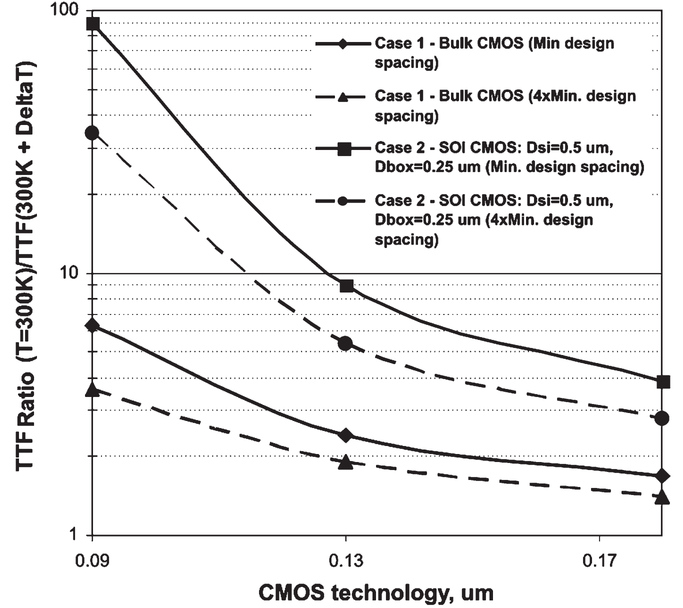

Fig. 16. Metallization lifetime degradation due to the self-heating effect versus technology scaling.

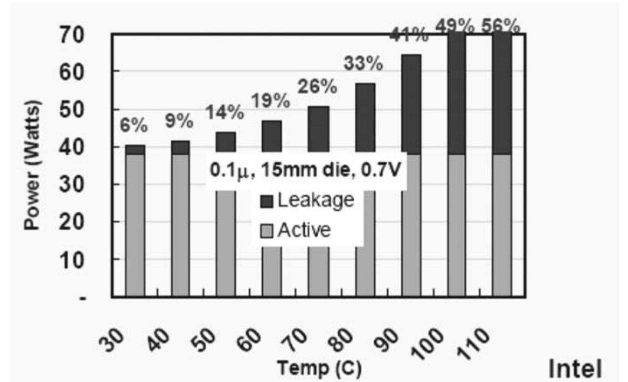

Fig. 17. Leakage and active power consumption of a 0.1-μm CMOS chip versus temperature, adopted from [2].

consumption. Assuming that the average current density (J_{ave}) in interconnects is increased with temperature similarly to the total power consumption of the chip, we can recalculate the TTF degradation for the 0.09-μm CMOS technology using the power consumption values from Fig. 17 and the junction temperature increase due to the self-heating effect from Fig. 14. The obtained results are summarized in Table II. The results presented in Table II show that the assumption of temperature independence of the average current density (J_{ave}) in interconnects with technology scaling may significantly underestimate the lifetime degradation of the interconnects in the sub-0.09-μm CMOS technologies, especially the SOI technology.

B. Effect of Self-Heating on TBD of Ultrathin Gate Oxides

The experimental measurements of TBD of ultrathin gate oxides with thicknesses less than 40 Å show that the conventional E and $(1/E)$ T_{BD} models cannot provide the necessary accuracy for calculation and prediction [32]. Hence, starting from the 0.18-μm CMOS technology (T_{OX} range is about 26–31 Å),

TABLE II
METALLIZATION LIFETIME DEGRADATION RATIO [TTF(300 K)/TTF (300 K + ΔT)] FOR THE 0.09-μm CMOS TECHNOLOGY

Bulk & SOI Tech.	Constant J_{ave}	$J_{ave}=f(T)$
Bulk CMOS (Min design spacing)	6.3	7.3
Bulk CMOS (4xMin design spacing)	3.6	3.8
SOI CMOS (Min design spacing)	90	176
SOI CMOS (4xMin design spacing)	34	61

a new $T_{\rm BD}$ model is proposed [32], [33]. Experiments show that the generation rate of stress-induced leakage current and the charge to breakdown $(Q_{\rm BD})$ in ultrathin oxides are controlled by the gate voltage rather than the electric field. This model (6) includes the gate oxide thickness $(T_{\rm OX})$ and the gate voltage (V_G) [33].

$$T_{\rm BD} = T_0 \cdot \exp\left[\gamma\left(\alpha \cdot T_{\rm OX} + \frac{E_a}{kT_j} - V_G\right)\right] \quad (6)$$

where γ is the acceleration factor, E_a is the activation energy, α is the oxide thickness acceleration factor, T_0 is a constant for a given technology, and T_j is the average junction temperature. TBD physical parameter values were extracted from experiments as follows: $(\gamma \cdot \alpha) = 2.01/\text{Å}$, $\gamma = 12.5\ 1/\text{V}$, and $(\gamma \cdot E_a) = 575$ meV [34]. We used (6) for the estimation of gate oxide TBD degradation due to the self-heating effect. In our calculations, T_j was equal to the peak junction temperature in high-performance clock drivers implemented in different CMOS technologies as shown in Fig. 9. The typical gate oxide thickness for the given CMOS technology can be adopted from the International Technology Roadmap for Semiconductors (ITRS) [27]. The obtained results are depicted in Fig. 18.

These results show that the $T_{\rm BD}$ degradation of gate oxides due to self-heating effect becomes significant for sub-0.09-μm CMOS technologies. The increased interfinger spacing in MOSFETs can reduce the $T_{\rm BD}$ degradation by 2.2× for the 0.09-μm PD SOI CMOS technology and by 1.6× for the 0.09-μm bulk CMOS technology. Another conclusion is that the self-heating effect has a stronger impact on the interconnect lifetime degradation than on the gate oxide degradation. For example, the interconnect lifetime degrades by $\sim 2\times$ in comparison with the degradation of the gate oxide TBD.

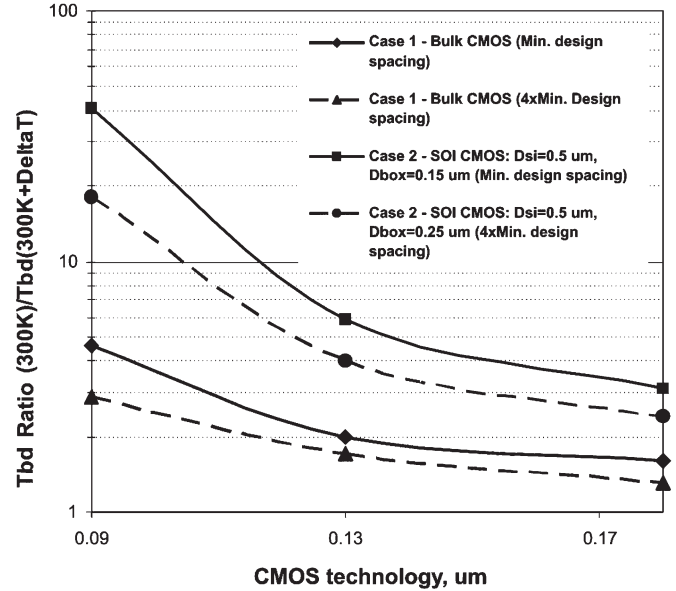

Fig. 18. TBD degradation due to the self-heating effect versus technology scaling.

VII. CONCLUSION

The self-heating effect in high-performance sub-0.18-μm CMOS circuits was investigated using fast transient and quasi-dc thermal simulations. We found that the 0.18-μm PD SOI technology has a 2.5× stronger self-heating effect ($\Delta T = 16$ K) than the 0.18-μm bulk CMOS technology ($\Delta T = 6$ K). The increase of the BOX thickness by 6× (thick BOX SOI technology is used for the smart-power applications), and the reduction of the silicon-film thickness by 17× (ultrathin silicon films are used for FD-SOI technologies) result in a stronger self-heating effect of 3.3× and 1.4×, respectively. With technology scaling, the self-heating effect becomes more significant and threatens the long-term reliability requirements for the sub-0.09-μm SOI CMOS technologies. However, the increase of the interfinger spacing in multifinger transistors can reduce the peak temperature to acceptable values. Moreover, an optimized clock-driver layout with an increased interfinger spacing can reduce the performance degradation due to the self-heating effect by 1.3×–1.4× in 0.09-μm bulk and SOI CMOS technologies. Our projections of long-term reliability degradation with technology scaling show that the metallization lifetime degradation and the gate oxide TBD reduction become extremely high in sub-0.09-μm bulk and SOI technologies. To prevent these effects, the design optimization of clock drivers is not enough and new cooling solutions and package and heat-sink-thermal-resistance reduction methods are required.

ACKNOWLEDGMENT

The authors would like to thank V. Axelrad and A. Shibkov of Sequoia Design Systems for providing the Technology Computer Aided Design tool and the discussion of obtained results.

REFERENCES

[1] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, no. 4, pp. 23–29, Jul./Aug. 1999.
[2] S. Rusu. (2001). "Trends and challenges in VLSI technology scaling toward 100 nm," in *Proc. Eur. Solid-State Circuits Conf. (ESSCIRC)*, Villach, Austria, pp. 23–25. [Online]. Available: http://www.imec.be/esscirc/esscirc2001/C01_Presentations/404.pdf
[3] S. Thompson, P. Packan, and M. Bohr. (1998). "MOS scaling: Transistor challenges for the 21st century," *Intel Technol. J.*, pp. 1–19, vol. 2, no. 3, 3rd Quarter. [Online]. Available: http://developer.intel.com/technology/itj/q31998/pdf/trans.pdf
[4] E. J. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down," *IBM J. Res. Develop.*, vol. 46, no. 2/3, pp. 169–180, Mar.–May 2002.

[5] N. Rinaldi, “Thermal analysis of solid-state devices and circuits: An analytical approach,” *Solid State Electron.*, vol. 44, no. 10, pp. 1789–1798, Oct. 2000.
[6] O. Semenov, A. Vassighi, M. Sachdev, A. Keshavarzi, and C. F. Hawkins, “Burn-in temperature projections for deep sub-micron technologies,” in *Proc. Int. Test Conf.*, Charlotte, NC, 2003, pp. 95–104.
[7] A. Vassighi, O. Semenov, M. Sachdev, A. Keshavarzi, and C. F. Hawkins, “CMOS IC technology scaling and its impact on burn-in,” *IEEE Trans. Device Mater. Rel.*, vol. 4, no. 2, pp. 208–221, Jun. 2004.
[8] K. Xiu and M. Ketchen, “Thermal modeling of a small extreme power density macro on a high power density microprocessor chip in the presence of realistic packaging and interconnect structures,” in *Proc. Electronic Components and Technology Conf.*, Las Vegas, NV, 2004, pp. 918–923.
[9] R. Viswanath, V. Wakharkar, A. Watwe, and V. Lebonheur. (2000, Aug.). “Thermal performance challenges from silicon to systems,” *Intel Technol. J.*, vol. 4 no. 3, pp. 1–16. [Online]. Available: http://developer.intel.com/technology/itj/q32000/pdf/thermal_perf.pdf
[10] P. Tadayon. (2000, Aug.). “Thermal challenges during microprocessor testing,” *Intel Technol. J.*, vol. 4, no. 3, pp. 1–8. [Online]. Available: http://developer.intel.com/technology/itj/q32000/pdf/thermal.pdf
[11] K. Banerjee and R. Mahajan, *Intel Development Forum*, 2002. [Online]. Available: ftp://download.intel.com/research/silicon/Thermals_press_IDF_0902.pdf
[12] I. Sauciuc, G. Chrysler, R. Mahajan, and M. Szleper, “Air-cooling extension—Performance limits for processor cooling applications,” in *Proc. IEEE Semiconductor Thermal Measurement and Management (Semi-Therm) Symp.*, San Jose, CA, 2003, pp. 74–81.
[13] S. K. Mathew *et al.*, “Sub-500 ps 64-b ALUs in 0.18-μm SOI/bulk CMOS: Design and scaling trends,” *IEEE J. Solid-State Circuits*, vol. 36, no. 11, pp. 1636–1646, Nov. 2001.
[14] J. D. Warnock *et al.*, “The circuit and physical design of the POWER4 microprocessor,” *IBM J. Res. Develop.*, vol. 46, no. 1, pp. 27–51, Jan. 2002.
[15] *SEQUOIA Device Designer User’s Guide*. (1998–2004). Woodside, CA: Sequoia Design Systems. [Online]. Available: http://www.sequoiadesignsystems.com
[16] P. E. Gronowski *et al.*, “High-performance microprocessors design,” *IEEE J. Solid-State Circuits*, vol. 33, no. 5, pp. 676–686, May 1998.
[17] C. J. Anderson *et al.*, “Physical design of a fourth-generation POWER GHz microprocessor,” in *Proc. Solid-State Circuits Conf.*, San Francisco, CA, 2001, pp. 232–233, p. 451.
[18] C. C. Liu, J. Zhang, A. K. Datta, and S. Tiwari, “Heating effects of clock drivers in bulk, SOI, and 3-D CMOS,” *IEEE Electron Device Lett.*, vol. 23, no. 12, pp. 716–718, Dec. 2002.
[19] K. Skadron, T. Abdelzaher, and M. R. Stan, “Control-theoretic techniques and thermal-RC modeling for accurate and localized dynamic thermal management,” in *Proc. Symp. High-Performance Computer Architecture*, Boston, MA, 2002, pp. 17–28.
[20] N. Rinaldi, “On the modeling of the transient thermal behavior of semiconductor devices,” *IEEE Trans. Electron Devices*, vol. 48, no. 12, pp. 2796–2802, Dec. 2001.
[21] *OriginPro 7.5, Scientific Graphing and Analysis Software*. [Online]. Available: www.originlab.com
[22] G. Krieger and P. D. Einzinberg, “Thermal analysis of ESD-related hot spots,” *IEEE Trans. Electron Devices*, vol. 35, no. 9, pp. 1553–1556, Sep. 1988.
[23] P. Raha, S. Ramaswamy, and E. Rosembaum, “Heat flow analysis for EOS/ESD protection device design in SOI technology,” *IEEE Trans. Electron Devices*, vol. 44, no. 3, pp. 464–471, Mar. 1997.
[24] D. L. Blackburn and A. R. Hefner, “Thermal components models for electro-thermal network simulation,” in *Proc. 9th IEEE Semiconductor Thermal Measurement and Management (Semi-Therm) Symp.*, Austin, TX, 1993, pp. 88–98.
[25] O. Paul and H. Baltes, “Thermal conductivity of CMOS materials for the optimization of microsensors,” *J. Micromech. Microeng.*, vol. 3, no. 3, pp. 110–112, Sep. 1993.
[26] K. E. Goodson *et al.*, “Annealing-temperature dependence of the thermal conductivity of LPCVD silicon-dioxide layers,” *IEEE Electron Device Lett.*, vol. 14, no. 10, pp. 490–492, Oct. 1993.
[27] *International Technology Roadmap for Semiconductors (ITRS)*, 2003 Edition. [Online]. Available: http://public.itrs.net
[28] *Pentium III Processor: Thermal Design Guidelines, Application Note*, [Online]. Available: http://www.intel.com/design/pentiumiii/applnots/24508702.pdf
[29] P. J. Restle and A. Deutsch, “Designing the best clock distribution network,” in *Proc. IEEE Symp. VLSI Circuits*, Honolulu, HI, 1998, pp. 2–5.
[30] B. A. Floyd and K. O. Kenneth, “The projected power consumption of a wireless clock distribution system and comparison to conventional distributing systems,” in *Proc. IEEE Int. Conf. Interconnect Technology*, San Francisco, CA, 1999, pp. 248–250.
[31] B. K. Liew, N. W. Cheung, and C. Hu, “Effects of self-heating on integrated circuit metallization lifetimes,” in *IEDM Tech. Dig.*, Washington, DC, 1989, pp. 323–326.
[32] J. H. Suehle, “Ultrathin gate oxide reliability: Physical models, statistics, and characterization,” *IEEE Trans. Electron Devices*, vol. 49, no. 6, pp. 958–971, Jun. 2002.
[33] P. E. Nicollian, W. R. Hunter, and J. C. Hu, “Experimental evidence for voltage driven breakdown models in ultra thin gate oxides,” in *Proc. IEEE Int. Reliability Physics Symp.*, San Jose, CA, 2000, pp. 7–15.
[34] F. Monsieur, E. Vincent, D. Roy, S. Bruyere, G. Pananakakis, and G. Ghibaudo, “Time to breakdown and voltage to breakdown modeling for ultra-thin oxides ($\mathrm{TOX} < 32$ Å),” in *Proc. IEEE Int. Reliability Workshop (IRW)*, Lake Tahoe, CA, 2001, pp. 20–25.

Oleg Semenov (M’05) received the Engineer degree (with honors) and the Ph.D. degree in microelectronics technology from the Moscow Institute of Electronics Engineering (Technical University), Moscow, Russia, in 1993 and 1996, respectively, and the M.Sc. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 2001. His Ph.D. research was on the investigation and development of silicon-on-insulator structures using a selective chemical etching of silicon.

From 1996 to 1998, he was with Korona Semiconductor, Moscow, where he worked as a Process Engineer. He is currently a Research Assistant Professor in the Department of Electrical and Computer Engineering, University of Waterloo. His research interests include reliability, testing, thermal management and manufacturing issues of deep submicrometer CMOS integrated circuits, impact of technology scaling on MOSFET characteristics, and design of electrostatic-discharge protection circuits. He is the author or coauthor of more than 20 papers in various international journals and conferences.

Arman Vassighi (S’01–A’04) received the B.S. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1990 and the M.S. and Ph.D. degrees from the University of Waterloo, Waterloo, ON, Canada, in 2000 and 2004, respectively.

He is currently with Intel Corporation, Hillsboro, OR. His research interests include very large scale integration (VLSI) low-power design and test and the quality and reliability of VLSIs at device, circuit, and system levels.

Manoj Sachdev (M’87–SM’97) received the B.E. degree (with honors) in electronics and communication engineering from the University of Roorkee, Roorkee, India and the Ph.D. degree from Brunel University, London, U.K., in 1984 and 1996, respectively.

From 1984 to 1989, he was with the Semiconductor Complex Ltd., Chandigarh, India, where he designed CMOS integrated circuits. From 1989 to 1992, he worked in the ASIC (Application-Specified Integrated Circuit) division of SGS–Thomson at Agrate, Milan, Italy. In 1992, he joined Philips Research Laboratories, Eindhoven, where he researched on various aspects of very large scale integration (VLSI) testing and manufacturing. He is currently a Professor in the Electrical and Computer Engineering Department, University of Waterloo, Waterloo, ON, Canada. His research interests include low-power high-performance digital circuit design, mixed-signal circuit design, and test and manufacturing issues of integrated circuits. He is the author or coauthor of 2 books, 2 book chapters, and 125 technical articles in conferences and journals. He is the holder or more than 15 granted and several pending U.S. patents in the broad area of VLSI circuit design and test.

Dr. Sachdev received several awards including the 1997 European Design and Test Conference Best Paper Award, the 1998 International Test Conference Honorable Mention Award, and the 2004 VLSI Test Syposium Best Panel Award.