Effect of CMOS Technology Scaling on Thermal Management During Burn-In

Oleg Semenov, Arman Vassighi, Manoj Sachdev, *Senior Member, IEEE,* Ali Keshavarzi, and C. F. Hawkins

***Abstract*—Burn-in is a quality improvement procedure challenged by the high leakage currents that are rapidly increasing with IC technology scaling. These currents are expected to increase even more under the new burn-in environments leading to higher junction temperatures, possible thermal runaway, and yield loss during burn-in. The authors estimate the increase in junction temperature with technology scaling. Their research shows that under normal operating conditions, the junction temperature is increasing 1.45×/generation. The increase in junction temperature under the burn-in condition was found to be exponential. The range of optimal burn-in voltage and temperature is reduced significantly with technology scaling.**

***Index Terms*—Burn-in testing, CMOS technology scaling, junction temperature, thermal management.**

I. Introduction and Motivation

THE ABILITY to improve performance with reduced power consumption per logic gate made CMOS the dominant technology for integrated circuits. Transistor scaling is the primary factor driving speed performance in microprocessors and memories. Historically, CMOS technology scaling per technology node has: 1) reduced gate delay by 30% allowing an increase in operating frequency of about 43%; 2) doubled transistor density; and 3) reduced energy per transition by about 65% while saving 50% of power [1]. To achieve this, transistor width, length, and oxide dimensions were scaled by 30%. As a result, the chip area decreased by 50% for the same number of transistors, and total parasitic capacitance decreased by 30%. Recent data on microprocessor operating frequencies show this trend [2]. Fig. 1 shows the evolution of Intel microprocessor operating clock frequency and gate delays per clock since 1987.

The supply voltage and transistor threshold voltages (V_{TH}) are also reduced by 30% under the constant electric field scaling scenario. V_{TH} must be scaled to maintain a sufficient gate overdrive $(V_{DD} - V_{TH})^n$ where n varies between 1 and 2 [3]. V_{TH} scaling has a erious impact on increased leakage current. Subthreshold leakage is an inverse exponential function of V_{TH}, so that the chip leakage power increases exponentially with technology scaling.

Manuscript received September 26, 2002; revised June 4, 2003.

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

A. Keshavarzi is with the Microprocessor Research Laboratories, Intel Corporation, Hillsboro, OR 97124-6497 USA.

C. F. Hawkins is with the Electrical and Computer Engineering Department, University of New Mexico, Albuquerque, NM 87131 USA.

Digital Object Identifier 10.1109/TSM.2003.818985

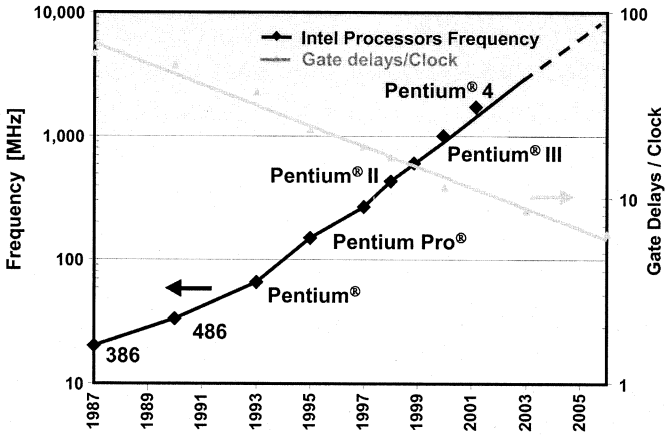


Fig. 1. Processor frequency trend adopted from [2].

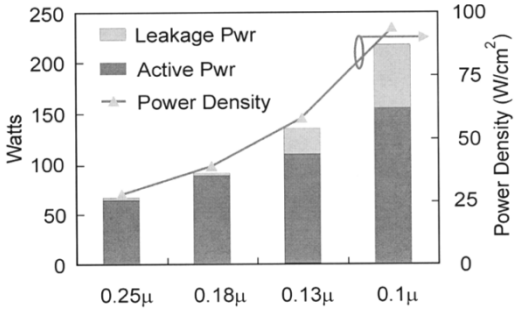


Fig. 2. Power density trend adopted from [2]. Assumptions: 15 mm die, 1.5× frequency increase per generation.

Scaling of technology results in higher transistor density and higher clock frequency. The increased clock frequency elevates the dynamic power. Historical trends also suggest that higher transistor density enables higher levels of functional integration and chip area. The total power consumption of high performance microprocessors increases, as Fig. 2 illustrates, for some Intel microprocessors. Note the increasing percentage of off-state leakage current at the 130- and 100-nm nodes.

The increased power consumption and higher clock frequency compromise long-term IC reliability. Techniques are needed for reliability estimation and prediction. The present methods include reliability simulation and accelerated laboratory tests. Reliability prediction is also linked to overall risk management, providing estimation of a reliable risk when a new technology is used without previous field experience.

A crucial parameter of reliability prediction procedures and burn-in testing is the average junction temperature. Junction

temperature is defined as the temperature of the Si lattice, and it has increased significantly due to the increased power consumption in high performance processors. For example, the measured junction temperature of a 1-GHz 64-bit RISC microprocessor implemented in 0.18-um CMOS technology was reported as 135 °C at $\mathrm{V_{DD}} = 1.9$ V [4]. This microprocessor had 15.2 million transistors packed in the 210 mm^2 chip area. Reliability prediction procedures and the estimation accuracy of optimal burn-in operating conditions ($\mathrm{V_{DD}}$, temperature, and time) decrease when transistor geometries are scaled, and we use new materials, technology processes, and operating environments.

There are several ways to estimate junction temperature. One method directly measures junction temperature via thermal sensors at several on-chip locations during normal and burn-in conditions [5], [6]. Another method uses chip-level three-dimensional (3-D)-electrothermal simulators that can find the steady-state CMOS very large-scale integration (VLSI) chip temperature profile at the corresponding circuit performance [7], [8]. Both methods have limitations. Thermal sensors are relatively large devices, and a number of them must be placed on the IC for accurate prediction. Furthermore, such sensors may require calibration. Gerosa *et al.* reported a thermal sensor with a sensing range of 0 °C–128 °C and a 5-bit resolution (4 °C). The size of each sensing element was 0.2 mm^2 [9]. Moreover, thermal sensors can only be used for verification. One may have to resort to other techniques for prediction and estimation. 3-D-electothermal simulators are one approach but cannot be used for large-scale integrated circuits such as microprocessors because of large CPU time. The simulation time of a two-dimensional (2-D) discrete cosine transformation (DCT) chip (107 832 transistors, 8 MHz) was reported at approximately 12 h [8].

The main objective of this paper is to propose a method for average junction temperature $(\mathrm{T_j})$ estimation under nominal and stressed conditions. The method can also predict the impact of scaling on average junction temperature. In this work, we focus on the intrinsic behavior of the die and do not consider the packaging issues, such as thermal impedance of the package and other considerations. Junction temperature estimation is crucial for burn-in condition optimization and reliability prediction.

The paper is organized as follows. Section II discusses the junction temperature as a parameter for reliability-prediction procedures. The thermal resistance models of transistors and their impact on the junction temperature are explained in Section III. The impact of CMOS technology scaling on average junction temperature increase at normal and burn-in conditions is analyzed in Section IV. In Section V, we discuss an optimization procedure for burn-in conditions to avoid thermal runaway.

II. Junction Temperature as a Parameter of Reliability Prediction Procedures

The effects of temperature on microelectronics devices are often assessed by accelerated tests carried out at high temperatures to generate reliability failures in a reasonable time period. Methods such as burn-in are often employed as reliability screens to weed out infant mortalities. Weak gate oxides are one of the major components of such failures. These failures are accelerated due to elevated temperature. There are several dielectric breakdown models available in the literature that can describe intrinsic as well as the defect related breakdown. In the following subsection, we consider three widely used models. It is apparent that junction temperature has influence in time to breakdown.

A. Time-Dependent Dielectric Breakdown Models (TDDB)—Gate Oxide Breakdown Models [10], [11]

The E and 1/E models are widely used in intrinsic gate oxide reliability predictions. Both models have physical basis. The E-model is expressed as

$$t = A\exp(-\gamma E)\exp\left(\frac{E_a}{kT_j}\right) \tag{1}$$

where t is the time to breakdown, A is a constant for a given technology, γ is the field acceleration parameter with unit as megavolts per centimeter, E is the oxide field, E_a is the thermal activation energy, k is Boltzmann's constant, T_j is the junction temperature (K). The E-model is based on thermo-chemical foundation. If we assume that the breakdown process is a current driven process, then the $1/E$ model predicts

$$t = \tau_0 \exp\left(\frac{G}{E}\right)\exp\left(\frac{E_a}{kT_j}\right) \tag{2}$$

where τ_0 and G are constants, E is the oxide field, E_a is the activation energy, and T_j is the junction temperature.

To increase the drive current and to control the short channel effects, the oxide thickness should decrease at each technology node. The experimental measurements of time to breakdown of ultrathin gate oxides with thickness less than 40 Å show that the conventional E and $1/E$ TDDB models cannot provide the necessary accuracy of calculation [12]. Hence, starting from 0.13-um CMOS technology (T_{OX} range is approximately 26–31 Å), a new TDDB models should be applied. An advanced TDDB models is the voltage driven breakdown (VDB) model [12], [13]. The experiments show that the generation rate of stress-induced leakage current (SILC) and charge to breakdown (Q_{BD}) in ultrathin oxides are controlled by gate voltage rather than electric field. Recently, a new time to breakdown model was proposed [14]. This model (3) includes the gate oxide thickness (T_{OX}) and the gate voltage (V_G)

$$T_{\mathrm{bd}} = T_0 \cdot \exp\left[\gamma\left(\alpha \cdot T_{\mathrm{OX}} + \frac{E_a}{kT_j} - V_G\right)\right] \tag{3}$$

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. Time to breakdown physical parameter values were extracted from experiments as follows: $(\gamma \cdot \alpha) = 2.01/\text{Å}, \gamma = 12.5$ 1/V, and $(\gamma \cdot E_a) = 575$ meV [14].

As mentioned before, all of the above methods describe the behavior of the intrinsic, good quality gate oxide. However, these models can also predict the time to breakdown under extrinsic oxide breakdown conditions, which include oxide damage by ion implantation, plasma damage, mechanical stresses, and contamination from technology processes. Under these conditions, the E_a is reduced. Since, the time to break-

down is a strong function of T_j and E_a, the above-mentioned oxide breakdown models can be used to predict the defect related breakdown.

B. Temperature and Voltage Acceleration Factor Models

Several industrial reliability standards are based on temperature and voltage acceleration factor models. The Mil-Hdbk-217F U.S. military standard defines the temperature acceleration factor as [15]

$$\pi_T = 0.1 \exp\left(-A\left(\frac{1}{T_j} - \frac{1}{298K}\right)\right) \tag{4}$$

where A is the constant and T_j is the junction temperature (K). The voltage acceleration factor is defined in the CNET reliability procedure as [16]

$$\pi_V = A_3 \exp\left[A_4 V_A \left(\frac{T_j}{298}\right)\right] \tag{5}$$

where A_3 and A_4 are constants, V_A is the applied voltage, and T_j is the junction temperature (K).

These reliability prediction models show that the average junction temperature is a fundamental parameter and should be accurately estimated for each technology generation. We must understand the properties of new materials and processes used for realizing VLSIs.

III. Thermal Resistance Models of Semiconductor Devices

The Arrhenius model predicts that the failure rate of integrated circuits is an inverse exponential function of the junction temperature. A small increase of 10 °C–15 °C in junction temperature may result in ~2× reduction in the lifespan of the device [17]. While T represents the ambient temperature for an IC, the relationship between ambient and average junction temperature for a VLSI is often described as in [18]

$$T_j = T + P \times R_{\text{th}} \tag{6}$$

where T is the ambient temperature, P is the total power dissipation of the chip, and R_{th} is the junction-to-ambient thermal resistance. One must analyze the impact of technology scaling on (6) to estimate the average junction temperature for several technologies. We investigated how the power dissipation and thermal resistance change with technology scaling.

Historically, the initial investigations on technology scaling and thermal resistance were carried out on bipolar transistors. For these devices, the thermal resistance was estimated as in [19]

$$R_{\text{th}} \approx \frac{1}{4K(L \times W)^{1/2}} \tag{7}$$

where K is the thermal conductivity of silicon, $(L \times W)$ is the emitter size, and R_{th} is the thermal resistance (°C per milliwatt). It was shown that the thermal resistance increased as the emitter size was reduced. Recently, a relationship between the thermal resistance of a MOSFET and its geometrical parameters was derived using a 3-D heat flow equation [20]. This equation is shown below and is obtained for bulk technologies where substrate thickness is significantly thicker than the thickness of device layer and the thermal impedance of the bulk is substantially smaller than that of the device

$$R_{\text{th}} = \frac{1}{2\pi K}\left[\frac{1}{L}\ln\left(\frac{L+\sqrt{W^2+L^2}}{-L+\sqrt{W^2+L^2}}\right) + \frac{1}{W}\ln\left(\frac{W+\sqrt{W^2+L^2}}{-W+\sqrt{W^2+L^2}}\right)\right] \tag{8}$$

where K is the thermal conductivity of silicon ($K = 1.5\times10^{-4}$ W/μm °C [21]) and L and W are channel geometry parameters. The thermal conductivity of silicon has a temperature dependence described as [22].

$$K = 154.86 \times (300/T)^{4/3} \quad (W{\cdot}m^{-1}{\cdot}K^{-1}). \tag{9}$$

However, in our calculations we assumed that the thermal resistance of silicon was temperature independent [20], [21]. The temperature dependence of silicon thermal conductivity is more important in silicon-on-insulator (SOI) technologies where self-heating contributes to a rise in junction temperature. We used the model of (8) for thermal resistance calculations for MOSFETs in different CMOS technologies.

TABLE I
CMOS Inverter Parameters and Performance Obtained From Simulations

CMOS Tech., um	N-MOSFET, W/L (um/um)	P-MOSFET, W/L (um/um)	N-MOSFET load, W/L (um/um)	F_{max}, MHz	$F_{operating} = 0.7 \times F_{max}$, MHz
0.35	4.0/0.35	10.0/0.35	3.0/3.5	1450	1015
0.25	2.86/0.25	7.14/0.25	2.15/2.5	1950	1365
0.18	2.06/0.18	5.14/0.18	1.55/1.8	2300	1610
0.13	1.49/0.13	3.71/0.13	1.12/1.3	4000	2800

IV. Scaling, Junction Temperature, and Normal and Burn-In Conditions

Rising junction temperature is a major issue for high performance circuits. In low-power applications, the power-supply voltage and transistor sizing are scaled more aggressively to minimize the power consumption [23], [24]. The transistor threshold voltage is typically higher than for high-performance ICs to suppress the subthreshold leakage. At the same time, the speed relative to the high-performance case should not degrade more than 1.5× [23]. We will focus on high-performance applications where dynamic and static power consumption are considerably high and pose a serious reliability threat.

We define $F_{\max}$ as the maximum toggle frequency of an inverter in a given technology. For dynamic power consumption calculation under normal operating conditions, we considered 70% of $F_{\max}$. HSPICE simulations were carried out with BSIM model level 49. Transistor models for 0.13-um CMOS technology were adopted from United Microelectronics Corporation (UMC). Transistor models for other CMOS technologies were adopted from Taiwan Semiconductor Manufacturing Corporation (TSMC). The simulation results and transistor sizes are given in Table I. The inverter's load was the standard load element (N-MOSFET) used by the TSMC for inverter ring-oscil-

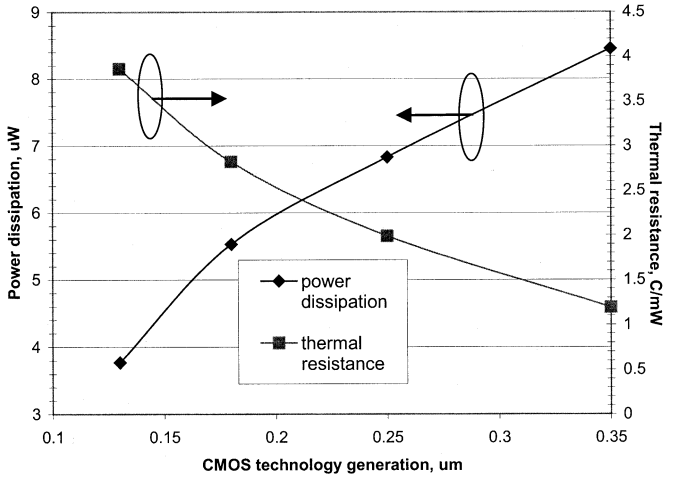


Fig. 3. Inverter power dissipation and transistor thermal resistance versus CMOS technology scaling.

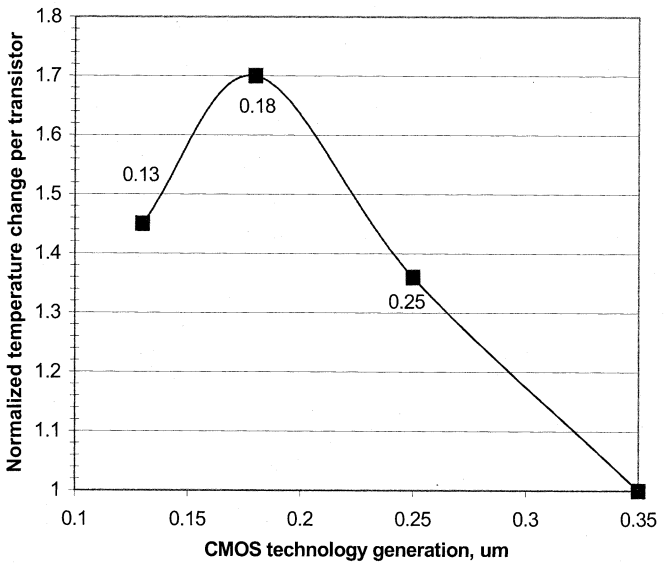


Fig. 4. Impact of technology scaling on junction temperature change in MOSFET.

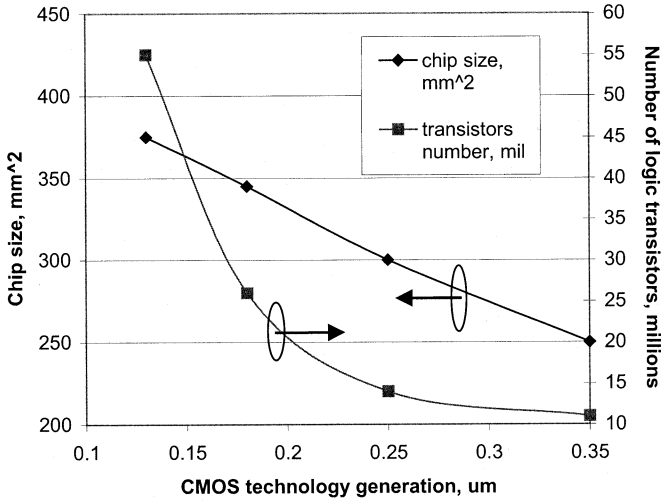


Fig. 5. Trend of CMOS logic chips progress (data for graphs were adopted from [25] and [26]).

lator simulations. The load element sizes were taken from the TSMC and UMC SPICE model file specified for each analyzed CMOS technology.

We simulated the total power consumption of an inverter toggling at $0.7F_{\max}$ in four different technologies. Thermal resistance of an average transistor was computed from (8). The average size of a transistor was achieved by averaging the n-MOS and p-MOS transistor widths. Since the transistor dimensions were reduced, the thermal resistance increased with scaling. Fig. 3 illustrates inverter power dissipation at operating frequency $0.7F_{\max}$ and thermal resistance of an average transistor as functions of technology.

The 0.35-um CMOS technology was used as the reference technology in this investigation. Equation (6) defines ΔT as the temperature difference between junction and the ambient.

If ΔT is defined as unity for 0.35-um technology, then we may calculate the normalized change in ΔT with respect to the reference technology. Using (6) and data presented in Fig. 3, we estimated the normalized average temperature increase for different technologies

$$\frac{\Delta T_{0.25-\mathrm{CMOS}}}{\Delta T_{0.35-\mathrm{CMOS}}} = \frac{(T_j - T)_{0.25-\mathrm{CMOS}}}{(T_j - T)_{0.35-\mathrm{CMOS}}} = \frac{(P \times R_{\mathrm{th}})_{0.25-\mathrm{CMOS}}}{(P \times R_{\mathrm{th}})_{0.35-\mathrm{CMOS}}}. \quad (10)$$

Fig. 4 shows the normalized MOSFET junction temperature change with respect to the 0.35-um technology using (10). As the technology went from 0.35 to 0.18 um, the normalized temperature increased primarily from the increase in thermal resistance with scaling. However, scaling from 0.18 to 0.13 um resulted in a lower normalized MOSFET junction temperature with respect to 0.18-um technology, since the power supply voltage was drastically reduced.

We must also consider the increase in transistor density with scaling when estimating the average normalized temperature increase. The density numbers were adopted from the International Technology Roadmap for Semiconductors (ITRS) [25], [26]. Fig. 5 illustrates the increased numbers of transistors and chip size with scaling. These graphs allow us to calculate the transistor density in the chip for the given technology.

The normalized temperature increase of a CMOS chip with technology scaling was calculated by multiplying the temperature increase per transistor in Fig. 4 and the transistor density calculated from Fig. 5. The results are shown in Fig. 6.

We conclude from Fig. 6 that the normalized temperature increase of the chip is elevated almost linearly with CMOS technology scaling from 0.25 to 0.13 um under normal operating conditions. The estimated junction temperature of a 0.13-um CMOS chip is $\sim$3.2 times higher than the junction temperature of 0.35-um CMOS chip. This calculation assumed that the ambient temperature was the same for all analyzed technologies.

A. Estimation of Junction Temperature Increase With Technology Scaling at Burn-In Conditions

The burn-in screening procedure weeds out latent defects from a product and thereby improves the outgoing quality and

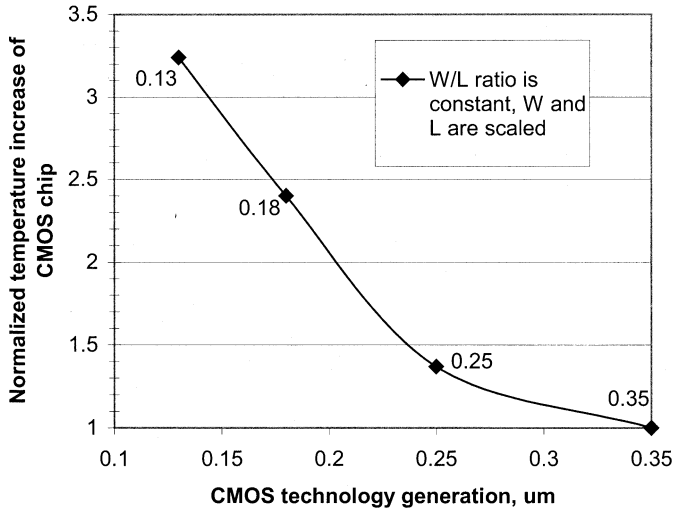


Fig. 6. Impact of technology scaling on normalized junction temperature increase of CMOS chip.

reliability of the product. During burn-in, ICs are subjected to elevated temperature and voltage in excess of normal operating conditions for a specific period of time. This accelerates the product lifetime through the early part of its life cycle allowing removal of the products that would have failed during that time.

There are die-level burn-in (DLBI) and wafer-level burn-in (WLBI) techniques. The DLBI can handle, contact, and do burn-in stress on several packaged die together, while WLBI has the ability to contact every die location and perform the burn-in test simultaneously on an entire wafer. We restricted ourselves to estimates of junction temperature in conventional static WLBI [27], [28]. For the DLBI, one must also consider the thermal impedance network of the package [29]. Once this network is known, then (6) can be suitably modified to reflect the total thermal resistance $(R_{\rm th})$ of the die and many types of packages.

We estimated the average power of inverters for different operating conditions and technologies by simulating the inverters at different temperatures and $V_{\rm DD}$. For static burn-in testing, we varied the stress temperature from 25 °C to 125 °C. Similarly, the stress voltage was varied from nominal $V_{\rm DD}$ for the given technology to $V_{\rm DD}+30\%$, and in this simulation (BSIM model level 49) the inverter input was grounded. The simulation gave $I_{\rm av}$, and the calculated values of P and ΔT are given in Table II. In this table, $I_{\rm av}$ and P are the average current and power dissipation of an inverter, and ΔT is $(T_j - T)$ per 1 mm^2 of chip area calculated using

$$\Delta T = P_{\rm transistor} \times R_{\rm th-transistor} \times \frac{D_{\rm density}}{2} \quad \left[\frac{^\circ{\rm C}}{{\rm mm}^2}\right] \tag{11}$$

where $P_{\rm transistor}$ is the power dissipation of the off-mode transistor in the inverter, $R_{\rm th-transistor}$ is the thermal resistance of the on-transistor in the inverter, and $D_{\rm density}$ is the transistor density in the CMOS chip. For a given technology, the thermal resistance was extracted from Fig. 3 and the transistor density was calculated from Fig. 5, respectively. We assumed a fully static CMOS design. Therefore, half of the total number of transistors are in the off-mode during static wafer level burn-in, and this was taken into account by dividing by 2 in (11).

Each off-mode transistor in a 1-mm^2 chip area was considered as an independent heat source, and the total junction temperature increase of this area over ambient temperature was defined as the multiplication of heat source density and the junction temperature increase of a single transistor. In practice, we must consider the thermal coupling effect of transistors on a chip, which depends on layout. In the first-order approximation, we neglected the thermal coupling effect of transistors in our analysis. Table II shows that the average leakage current and dissipated power is increased by at least two orders of magnitude by technology scaling if the ambient temperature is 85 °C or less, and at 125 °C, the increase in current and power dissipation with technology scaling is relatively less. However, the increase in ΔT is more dramatic due to increased transistor density, leakage current, and the thermal resistance. The normalized temperature increase of a CMOS chip with technology scaling at static wafer level burn-in conditions is shown in Fig. 7. For static burn-in conditions, Fig. 7 shows that the estimated average junction temperature increase in CMOS chip should be $\sim$70 times higher for 0.13-um technology than for the 0.35-um technology. This junction temperature increase with technology scaling is the result of a drastic stand-by leakage power increase, the higher transistor density in advanced CMOS die, and the thermal resistance increase of scaled MOSFETs.

V. Burn-In Limitations and Optimization to Avoid Thermal Runaway

Since there are major reliability failure mechanisms that are accelerated by temperature, burn-in testing is done at an elevated temperature. These mechanisms include metal stress voiding and electromigration, metal slivers bridging shorts, and gate-oxide wearout and breakdown [30]. However, there are physical and burn-in equipment related limitations for temperature and voltage stress. Die failure rate (failures per million) increases exponentially with temperature for most failure mechanisms [31]. As a result, there is a risk of increasing the yield loss if the burn-in conditions are overstressed. Hence, we should optimize the junction temperature of die for normal and burn-in conditions.

A. Physical Limits of Junction Temperature

The maximum operating temperatures for semiconductor devices can be estimated from semiconductor intrinsic carrier density, which depends on the band-gap of the material. When the intrinsic carrier density reaches the doping level of the active region of devices, electrical parameters are expected to change drastically. The highest operating junction temperature for standard silicon technology is about 200 °C; however, the circuit performance is reduced substantially [32]. The influence of temperature on some important MOSFET parameters is summarized in Table III.

The junction temperature of a PowerPC microprocessor implemented in a 0.35-um CMOS technology with a 0.3-um effective transistor channel lengths is about 90 °C–100 °C at an

TABLE II
DC SIMULATION (I_{av}) AND CALCULATION RESULTS ($P, \Delta T$) OF CMOS INVERTERS FOR DIFFERENT TECHNOLOGIES

CMOS technology		25 °C			85 °C			125 °C		
		I_{av}, nA	P, nW	ΔT, °C/mm²	I_{av}, nA	P, nW	ΔT, °C/mm²	I_{av}, nA	P, nW	ΔT, °C/mm²
0.35-um	V_{DD} = 3.3 V	0.0077	0.025	0.00071	0.07	0.23	0.0066	2.05	6.77	0.2
	V_{DD} = 3.8 V	0.0092	0.035	0.00099	0.084	0.32	0.0091	2.15	8.17	0.23
	V_{DD} = 4.3 V	0.0111	0.0477	0.0014	0.11	0.47	0.014	2.27	9.76	0.28
0.25-um	V_{DD} = 2.5 V	0.0193	0.0483	0.0023	0.418	1.04	0.05	3.96	9.9	0.29
	V_{DD} = 2.9 V	0.022	0.0638	0.0031	0.47	1.36	0.065	4.41	12.80	0.35
	V_{DD} = 3.25 V	0.025	0.0813	0.0039	0.531	1.75	0.08	4.81	15.87	0.45
0.18-um	V_{DD} = 1.8 V	0.0905	0.163	0.02	1.33	2.39	0.24	8.96	16.13	0.97
	V_{DD} = 2.1 V	0.101	0.21	0.022	1.48	3.08	0.31	9.75	20.48	1.23
	V_{DD} = 2.35 V	0.112	0.264	0.027	1.62	3.81	0.39	10.9	25.6	1.51
0.13-um	V_{DD} = 1.2 V	0.766	0.92	0.2	8.45	10	2.32	28	34	7.79
	V_{DD} = 1.4 V	1.20	1.68	0.38	12.3	17	3.94	34	47	10.97
	V_{DD} = 1.56 V	1.86	2.9	0.67	17.7	27.6	6.4	55	85	19.81

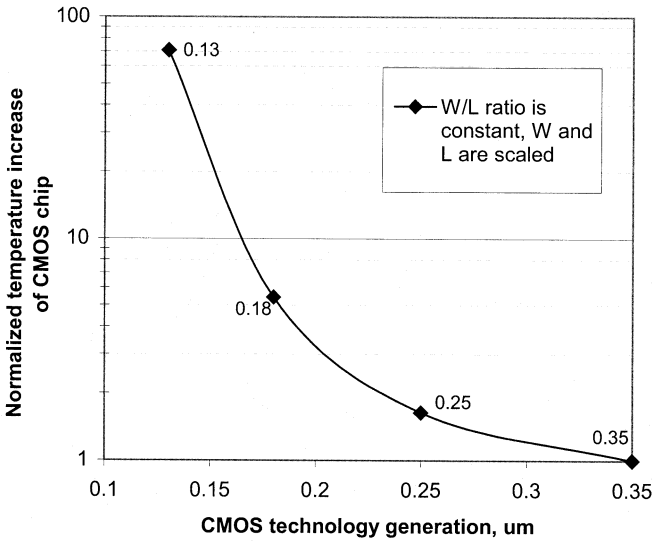


Fig. 7. Normalized junction temperature increase of CMOS logic chip at burn-in conditions ($V_{DD} + 30\%, T = 125$ °C).

TABLE III
TEMPERATURE DEPENDENCE OF IMPORTANT Si-MOSFET PARAMETERS (DATA ADOPTED FROM [32])

Parameter	Temperature dependence	Affected property
Thermal conductivity, K	$\approx T^{-1.6}$	Self heating
Built-in potential, V_{bi}	$kT/q \ln(N_A N_D/n_i(T)^2)$	~ +20% per 100 K
Threshold voltage, V_{TH}	$2\psi_B(T) + (4\varepsilon_{Si} q N_A \psi_B(T)/C_I)^{0.5}$	~ -0.8 mV/K
pn junction reverse current	$a\, n_i^2(T) + b\, n_i(T)/\tau_{sc}$	~ $+10^2$ to + 10^4 per 100 K

operating speed of 200–250 MHz [33], [9]. If we use this as the die reference temperature and assume that Fig. 6 estimates the junction temperature increase over room temperature (ΔT) with reasonable accuracy, then we should expect a 2.4× increase in junction temperature over room temperature for the same microprocessor implemented in a 0.18-um CMOS technology. Hence, the ΔT should be approximately 156 °C–180°C, provided that cooling techniques remain the same. Since this temperature is closed to the physical limit of junction temperature for silicon technology (200 °C), the advanced cooling techniques must be developed to reduce the junction temperature. Every second generation of scaling requires new cooling techniques to keep the junction temperature at an acceptable operational level (∼100 ° C).

B. Power Limitation of Burn-In Equipment

The total number of die that can be simultaneously powered up for burn-in testing will likely be limited by the maximum power dissipation capacity of the burn-in oven. A typical wafer may contain several hundred die. If all dies are active, then the total power dissipation can reach the several kilowatt range. Typically, burn-in ovens have a maximum dissipation power between 2500–6500 W [34]. If we use the power dissipation of a single transistor in an inverter at static stressed conditions from

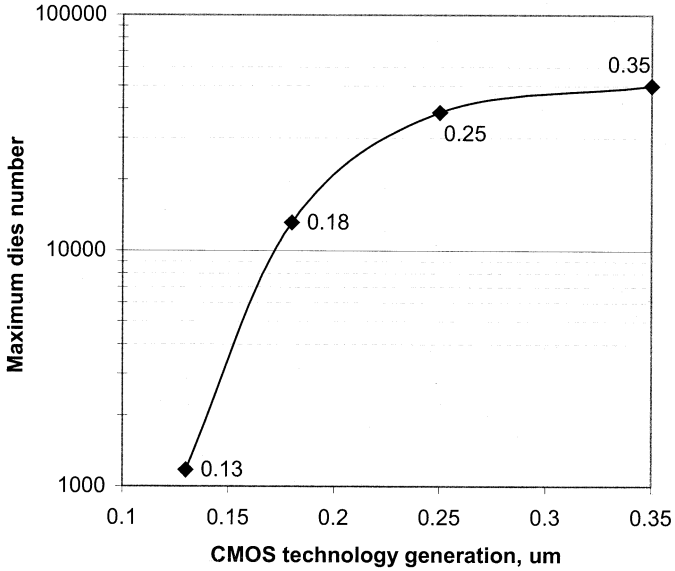


Fig. 8. Maximum die number for one burn-in load versus CMOS technology scaling. Maximum power dissipation of burn-in oven is 2500 W.

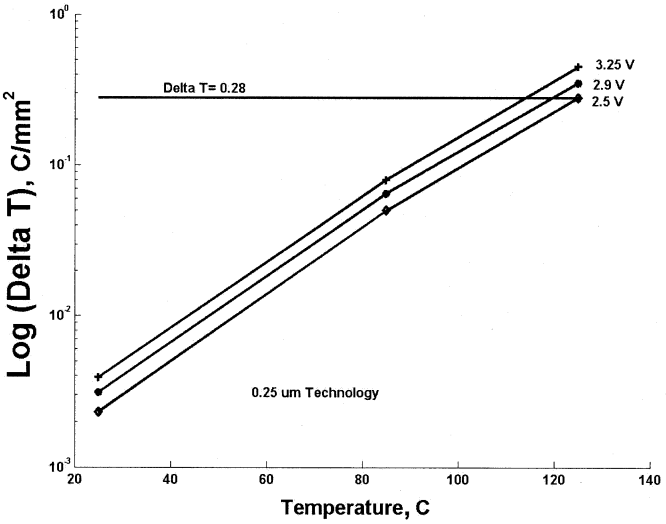


Fig. 9. Junction temperature increase over ambient stressed temperature per 1 mm^2 chip area versus stressed temperature and different $V_{\rm DD}$.

Table II, and the number of transistors of the logic chip for different CMOS technologies from Fig. 5, then we can estimate the maximum number of die for different technologies that can be simultaneously powered in a burn-in oven using

$$N_{\rm dies} = \frac{P_{\rm oven}}{P_{\rm transistor} \times \frac{N_{\rm transistors}}{2}} \tag{12}$$

where $P_{\rm oven}$ is the maximum power dissipation of the burn-in oven at stressed conditions, $P_{\rm transistor}$ is the power dissipation of a single transistor at static stressed conditions for the given technology, and $N_{\rm transistors}$ is the total number of transistors in the logic chip for the given technology. Equation (12) assumes that 50% of the total number of transistors are in off-mode at static burn-in testing (fully static CMOS design of the chip).

A typical burn-in oven, such as the PBC1-80 of Despatch Industries, has a maximum power dissipation of about 2500 W at 125 °C [34]. The room ambient temperature is assumed to be 25 °C. Now, from (12), we can calculate the maximum number of die that can be powered during burn-in. Fig. 8 plots this calculation over several technology generations. Fig. 8 shows that the maximum number of die that can simultaneously be powered in burn-in is exponentially reduced. The exponential increase in the standby power dissipation is the main cause for such a behavior.

C. Optimization of Burn-In Stress Conditions With Technology Scaling for Fixed Yield Loss

Burn-in yield is the fraction of the total number of stressed devices that meet nominal functional specifications after the burn-in. The optimal burn-in conditions for maintaining the projected failure rate requires that the defect distribution models and their growth models be studied. Although the burn-in is related to the removal of infant mortality, it may affect the yield of semiconductor devices. This is due to the fact that defects will grow during burn-in and some of them will cause yield loss. The amount of the defect growth and yield loss depends on burn-in environment, such as stressed voltage, stressed temperature, and burn-in time. The post burn-in yield loss was studied [35], [36], and Kim *et al.* proposed a post burn-in yield loss model [35]

$$Y_{\rm loss} = Y\left(1 - Y^{\frac{\upsilon}{1-\upsilon}}\right) \tag{13}$$

where Y is the yield loss before burn-in and υ is a constant that depends on stressed temperature and voltage. Using the $1/E$ gate oxide breakdown model (2) and the post burn-in yield loss model (13), Vassighi *et al.* demonstrated that the post burn-in yield loss is increased exponentially with the elevation of stressed temperature for the given stressed voltage [36]. This result was obtained for 0.18-um CMOS technology ($T_{\rm OX} \approx 41$ Å).

Hence, an overstressed die during burn-in may significantly increase the post burn-in yield loss, especially when junction temperatures at burn-in and normal operating conditions are increased with technology scaling. Thus, to a first order, we want a constant yield loss during burn-in testing with technology scaling. Burn-in temperature and voltage should be optimized for different CMOS technologies to maintain the average junction temperature of the die at the fixed level.

If electrical defect densities are equal, then we assume that the post burn-in yield loss for an advanced CMOS technology should not be worse than the post burn-in yield loss for the 0.35-um CMOS technology. This means that the junction temperature increase over ambient temperature during burn-in testing for advanced technologies should not be higher than the burn-in junction temperature increase for 0.35-um CMOS technology. From Table II, for a 0.35-um CMOS technology, the junction temperature increase (ΔT) over ambient stressed temperature per 1 mm^2 of chip is 0.28 °C at $V_{\rm DD} = 4.3$ V and $T = 125$ ° C. If we plot ΔT versus stressed temperature for different stressed voltages, we can find the optimal burn-in temperature and voltage when $\Delta T = 0.28$ ° C/mm^2 for other CMOS technologies. For example, Fig. 9 presents this technique for a 0.25-um CMOS technology using the data from Table II. Similarly, we can find the optimal burn-in temperature for other technologies using data from Table II. The results are shown in Fig. 10, where the optimal burn-in temperature is presented for different technologies. It is assumed in Fig. 10

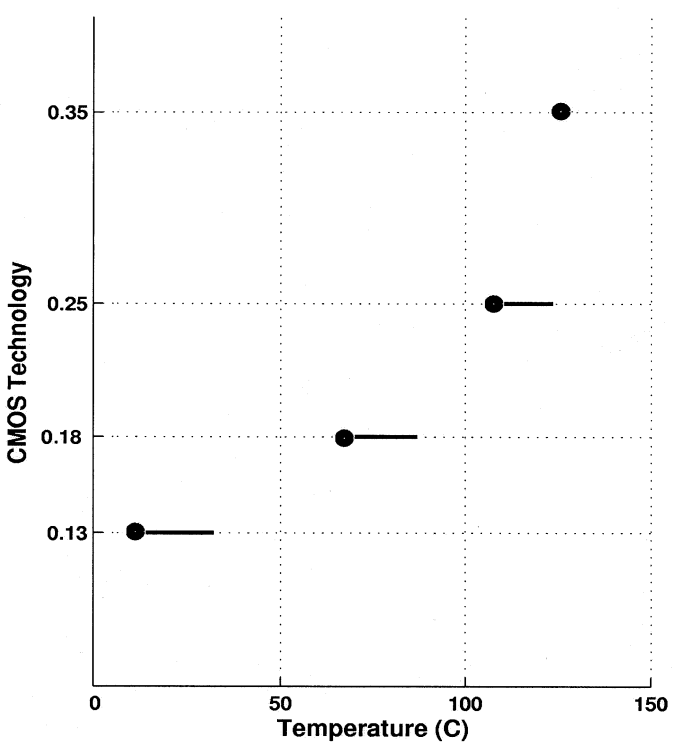


Fig. 10. Optimized burn-in voltage and temperature for constant burn-in loss.

that the junction temperature with the scaling remains constant. Since the power densities and device thermal impedance is increasing with scaling, the ambient temperature (burn-in) must be reduced with scaling, as shown in Fig. 10. In this case, we expect that the post burn-in yield loss for scaled CMOS technologies has the same value as the post burn-in yield loss for 0.35-um CMOS technology if the incoming electrical defect densities are the same for each technology.

Fig. 10 shows the optimal stressed temperature is significantly reduced with technology scaling. Recently presented data show that the optimal burn-in temperature used for microprocessors implemented in 0.18-um CMOS technology is 90 °C [37]. The expected optimal burn-in conditions for 0.13-um CMOS technology are $V_{\rm DD} \approx 1.4$ V and $T_{\rm burn-in} \approx 35°$ C. If such a trend continues, we will have to cool future generations of CMOS devices during the burn-in below room temperature, if we do not want the post burn-in loss worse than that of the 0.35-um CMOS technology. On the other hand, practical constraints (such as preventing condensation over long time periods) may not allow a drastic reduction in burn-in temperature. In such situations, the post burn-in loss may become larger.

As we scale the technology estimation of burn-in the temperature will become crucial. Therefore, a procedure must be evolved to accurately estimate this temperature. One such procedure is described in the Appendix.

VI. Conclusion

We investigated the impact of technology scaling on the burn-in environment and from the average junction temperature estimation, the following conclusions were obtained.

1) Our research shows a steady increase in junction temperature with scaling. Under normal operating conditions, the normalized increase in junction temperature is estimated as 1.45×/generation. The normalized junction temperature increase under the static burn-in conditions becomes exponential with technology scaling.
2) The number of die that can be simultaneously burnt-in is reduced exponentially with the technology scaling, because of the maximum power dissipation limit of burn-in ovens.
3) Finally, our research shows that the optimal stressed temperature in a burn-in environment is significantly reduced with technology scaling.

Appendix
Procedure for Average Junction Temperature Estimation

The procedure includes the following steps.

1) Calculate the thermal resistance of the average size MOSFET for the given technology and design using (7). For example, the average transistor size may be adopted from the SPICE technology file for a ring-oscillator circuit, or the average transistor size may be computed from a given design.
2) Estimate the average power dissipation of the average CMOS inverter at nominal operating conditions and operating frequency. Repeat it for dc and/or ac burn-in conditions.
3) Using (5) and the average transistor density for a given design, compute the average junction temperature of the chip. The typical transistors density for memories and microprocessors may be adopted from the International Technology Roadmap for Semiconductors (ITRS) [26].
4) Using (10), compute the difference between junction and ambient temperature (ΔT) under burn-in conditions.

This method may be used for prediction of optimal burn-in stressed conditions and average die temperature increase at nominal operating conditions, when CMOS technology is aggressively scaled down.

References

[1] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, pp. 23–29, July–Aug. 1999.
[2] S. Rusu. Trends and challenges in VLSI technology scaling toward 100 nm. presented at ESSCIRC. [Online]. Available: http://www.ess-circ.org/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.* [Online], pp. 1–19. Available: http://developer.intel.com/technology/itj/archive.htm
[4] J. Ahn, H.-S. Kim, T.-J. Kim, H.-H. Shin, Y.-H. Kim, D.-U. Lim, J. Kim, U. Chung, S.-C. Lee, and K.-P. Suh, "1 GHz microprocessor integration with high performance transistor and low RC delay," in *IEDM Tech. Dig.*, 1999, pp. 28.5.1–28.5.4.
[5] T. J. Goh, A. N. Amir, C.-P. Chiu, and J. Torresola, "Novel thermal validation metrology based on nonuniform power distribution for Pentium III Xeon cartridge processor design with integrated level two cache," in *Proc. Electronic Components and Technology Conf.*, 2001, pp. 1181–1186.
[6] S. H. Gunter, F. Binns, D. M. Carmean, and J. C. Hall. (2001) Managing the impact of increasing microprocessor power consumption. *Intel Technol. J.* [Online], pp. 1–9. Available: http://developer.intel.com/technology/itj/archive.htm

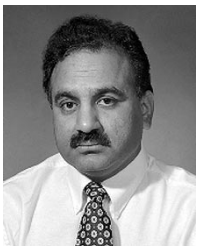
[7] Y. K. Cheng, C.-C. Teng, A. Dharchoudhury, E. Rosenbaum, and S.-M. Kang, "A chip-level electrothermal simulator for temperature profile estimation of CMOS VLSI chips," in *Proc. Int. Symp. Circuit Systems*, 1996, pp. 580–583.
[8] C.-C. Teng, Y.-K. Cheng, E. Rosenbaum, and S.-M. Kang, "iTEM: A temperature-dependent electromigration reliability diagnosis tool," *IEEE Trans. Computer-Aided Design* , vol. 16, pp. 882–893, Aug. 1997.
[9] P. Reed, M. Alexander, J. Alvarez, M. Brauer, C.-C. Chao, C. Croxton, L. Eisen, T. Le, T. Ngo, C. Nicoletta, H. Sanchez, S. Taylor, N. Vanderschaaf, and G. Gerosa, "A 250-MHz 5-W PowerPC microprocessor with on-chip L2 cash controller," *IEEE J. Solid-State Circuits*, vol. 32, pp. 1635–1649, Nov. 1997.
[10] J. W. McPherson, V. K. Reddy, and H. C. Mogul, "Field-enhanced Si-Si bond-breakage mechanism for time-dependent dielectric break-down in thin-film SiO_2 dielectrics," *Appl. Phys. Lett.*, vol. 71, no. 8, pp. 1101–1103, 1997.
[11] A. M. Yassine, H. E. Nariman, M. McBride, M. Uzer, and K. R. Olasupo, "Time dependent breakdown of ultra-thin gate oxide," *IEEE Trans. Electron Devices*, vol. 47, pp. 1416–1420, July 2000.
[12] J. H. Suehle, "Ultra thin gate oxide reliability: Physical models, statistics, and characterization," *IEEE Trans. Electron Devices*, vol. 49, pp. 958–971, June 2002.
[13] P. E. Nicollian, W. R. Hunter, and J. C. Hu, "Experimental evidence for voltage driven breakdown models in ultra thin gate oxides," *Proc. IEEE Int. Reliability Physics Symp.*, pp. 7–15, 2000.
[14] 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 $(\mathrm{T_{OX}} < 32\text{Å})$," *Proc. IEEE Int. Reliability Workshop (IRW)*, pp. 20–25, 2001.
[15] P. Lall, "Tutorial: Temperature as an input to microelectronics-reliability models," *IEEE Trans. Reliability*, vol. 45, pp. 3–9, Jan. 1996.
[16] J. B. Bowles, "A survey of reliability-prediction procedures for microelectronics devices," *IEEE Trans. Reliability*, vol. 41, pp. 2–12, Jan. 1992.
[17] R. Viswanath, V. Wakharkar, A. Watwe, and V. Lebonheur. (2000) Thermal performance challenges from silicon to systems. *Intel Technol. J.* [Online], pp. 1–16. Available: http://developer.intel.com/technology/itj/archive.htm
[18] P. Tadayon. (2000) Thermal challenges during microprocessor testing. *Intel Technol. J.* [Online], pp. 1–8. Available: http://developer.intel.com/technology/itj/archive.htm
[19] R. C. Joy and E. S. Schlig, "Thermal properties of very fast transistors," *IEEE Trans. Electron Devices*, vol. ED-17, pp. 586–594, Aug. 1970.
[20] N. Rinaldi, "Thermal analysis of solid-state devices and circuits: An analytical approach," *Solid-State Electron.*, vol. 44, pp. 1789–1798, 2000.
[21] ——, "On the modeling of the transient thermal behavior of semiconductor devices," *IEEE Trans. Electron Devices*, vol. 48, pp. 2796–2802, Dec. 2001.
[22] D. L. Blackburn and A. R. Hefner, "Thermal components models for electro-thermal network simulation," *Proc. 9th IEEE Semi-Therm Symp.*, pp. 88–98, 1993.
[23] B. Davari, R. H. Dennard, and G. G. Shahidi, "CMOS scaling for high performance and low power—The next ten years," *Proc. IEEE*, vol. 83, pp. 595–606, Apr. 1995.
[24] M. Borah, R. M. Owens, and M. J. Irwin, "Transistor sizing for low power CMOS circuits," *IEEE Trans. Computer-Aided Design*, vol. 15, pp. 665–671, June 1996.
[25] D. P. Valett and J. M. Soden, "Finding fault with deep-submicron ICs," *IEEE Spectrum*, pp. 39–50, Oct. 1997.
[26] International Technology Roadmap for Semiconductors [Online]. Available: http://public.itrs.net
[27] L. W. Ivy, Jr., P. Godavarti, N. Alizy, T. Mckenzie, and D. Mitchell, "Sacrificial metal wafer level burn-in KGD," *Proc. IEEE Electronic Components Technology Conf.*, pp. 535–540, 2000.
[28] J. Novitsky and D. Pedersen, "Form factor introduces an integrated process for wafer-level packaging, burn-in test, and module level assembly," *Proc. IEEE Int. Symp. Advanced Packaging Materials*, pp. 226–231, 1999.
[29] G. Kromann, "Thermal management of a C4/CBGA interconnect technology for a high-performance RISC microprocessor: The Motorola PowerPC 620™ microprocessor," *Proc. IEEE Electronic Technology Conf.*, pp. 652–659, 1996.
[30] A. W. Righter, C. F. Hawkins, J. M. Soden, and P. Maxwell, "CMOS IC reliability indicators and burn-in economics," in *Proc. Int. Test Conf.*, 1998, pp. 194–203.
[31] N. F. Dean and A. Gupta, "Characterization of a thermal interface material for burn-in application," *Proc. IEEE Thermal Thermomechanical Phenomena Electronic Syst.*, pp. 36–41, 2000.
[32] W. Wondrak, "Physical limits and lifetime limitations of semiconductor devices at high temperature," *Microelectronics Reliability*, vol. 39, no. 6–7, pp. 1113–1120, 1999.
[33] H. Sanchez, B. Kuttanna, T. Olson, M. Alexander, G. Gerosa, R. Philip, and J. Alvarez, "Thermal management system for high performance PowerPC microprocessors," *Proc. IEEE COMPCON*, pp. 325–330, 1997.
[34] [Online]. Available: http://www.despatch.com/pdfs/PBC.pdf
[35] T. Kim, W. Kuo, and W.-T. K. Chien, "Burn-in effect on yield," *IEEE Trans. Electron. Packaging Manuf.*, vol. 23, no. 4, pp. 293–299, 2000.
[36] A. Vassighi, O. Semenov, and M. Sachdev, "Impact of power dissipation on burn-in test environment for sub-micron technologies," *Proc. IEEE Int. Workshop Yield Optimization and Test*, 2001.
[37] T. M. Mak, "Is CMOS more reliable with scaling?," presented at the *IEEE Int. On-Line Testing Workshop*, July 2002.

Oleg Semenov received the Engineer degree (with honors) and Doctor of Science degree in microelectronics technology from the Moscow Institute of Electronics Engineering (Technical University), Russia, in 1993 and 1996, respectively, and the M.Sc. degree in electrical engineering from the University of Waterloo, ON, Canada, in 2001.

He was with Joint Stock Company (Hong Kong-Russia), Moscow, from 1996 to 1998 where he worked as a Process Engineer. Currently, he is a Postdoctoral Fellow in the Electrical and Computer Engineering Department, University of Waterloo, Canada. His research interests include reliability, testing and manufacturing issues of deep submicron CMOS ICs, the impact of technology scaling on MOSFET characteristics, and design of ESD protection circuits.

Arman Vassighi received the B.S. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1990, and the M.S. degree from University of Waterloo, ON, Canada, in 2000. He is currently working toward the Ph.D. degree at the University of Waterloo.

His research interests include VLSI low power and burn-in test optimization. His main focus is using device level and circuit level techniques to reduce off current of deep-submicron MOSFET while optimizing burn-in conditions.



Manoj Sachdev (SM'97) received the B.E. degree (with honors) in electronics and communication engineering from University of Roorkee, India, and the Ph.D. degree from Brunel University, U.K.

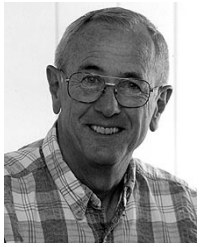
He was with Semiconductor Complex Limited, Chandigarh, India, from 1984 to 1989, where he designed CMOS integrated circuits. From 1989 to 1992, he worked in the ASIC division of SGS-Thomson at Agrate, Milan, Italy. In 1992, he joined Philips Research Laboratories, Eindhoven, where he researched various aspects of VLSI testing and manufacturing. He is currently a Professor in the Electrical and Computer Engineering Department, University of Waterloo, ON, Canada. His research interests include low-power and high-performance digital circuit design, mixed-signal circuit design, test and manufacturing issues of integrated circuits. He has written a book and two book chapters on testing of integrated circuits. He has contributed to more than 80 papers in various conferences and journals.

Dr. Sachcev received the best paper award for his paper at the European Design and Test Conference, 1997, and an honorable mention award for his paper at the International Test Conference, 1998. He holds more than ten granted and several pending U.S. patents in the area of VLSI design and test.

Ali Keshavarzi received the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN.

He is a Staff Research Scientist at Microprocessor Research Laboratories (MRL), Intel Corporation, Portland, OR. His current research interests include long-term research in low-power/high-performance circuit techniques and transistor device structures for future generations of microprocessors. He has been with Intel for 12 years, has published more than 20 papers, and has more than 20 patents (ten issued and the rest are pending patents).

Dr. Keshavarzi received the best paper award at 1997 IEEE International Test Conference at Washington, DC, on testing solutions of intrinsically leaky integrated circuits. He is a member of the ISLPED and ISQED technical program committees.



Chuck F. Hawkins is a Professor in the Electrical Engineering and Computer Engineering Department, University of New Mexico, where he teaches and does research in CMOS electronics, test, reliability, and failure analysis. He has worked with Sandia National Labs. IC Development Group since 1984 and has been a consultant with Intel, Philips Research Labs., and AMD Corp. He coauthored three books including (with J. Segura) *CMOS Electronics: How it Works, How it Fails* (New York: IEEE, 2003).

Dr. Hawkins is the Editor of the *Electron Device Failure Analysis* magazine and the past General and Program Chair of the International Test Conference. With coworkers at Sandia National Laboratories he has won several Best and Honorable mention papers at ITC and the International Symposium on Test and Failure Analysis (ISTFA).