CMOS IC Technology Scaling and Its Impact on Burn-In

Arman Vassighi, Oleg Semenov, Manoj Sachdev, *Senior Member, IEEE*, Ali Keshavarzi, and Chuck Hawkins

Invited Paper

***Abstract*—This article describes how CMOS IC technology scaling impacts semiconductor burn-in and burn-in procedures. 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 of good parts during burn-in. The paper discusses the effect of junction temperature on device reliability, aging, and burn-in procedure optimization. The effect of device thermal runaway and the requirements it forces on commercial burn-in ovens, device package, and device cooling are also described.**

***Index Terms*—Burn-in, junction temperature, packaging, reliability, technology scaling, thermal management.**

I. Introduction

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

Power supply voltage in scaled technologies must be lowered for two main reasons: 1) to reduce the device internal electric fields and 2) to reduce active power consumption since it is proportional to $V_{\mathrm{DD}}{}^2$. As V_{DD} scales, then V_{T} must also be scaled to maintain drain current overdrive $(V_{\mathrm{DD}}-V_{\mathrm{T}})$ to achieve higher performance. This lower V_{T} leads to higher off-state leakage current, and this is the major problem facing burn-in and scaled nanometer technologies.

The total power consumption of high-performance microprocessors increases with scaling. Off-state leakage current is an increasing percentage of the total current at the 130-nm and sub-100-nm nodes under nominal conditions. The ratio of leakage to active power becomes adverse under burn-in conditions and the off-state leakage is the dominant power. Typically, clock frequencies are kept in the tens of megahertz range during burn-in resulting in a substantial reduction in active power. On the other hand, the voltage and temperature stresses cause the off-state leakage to be the dominant power component.

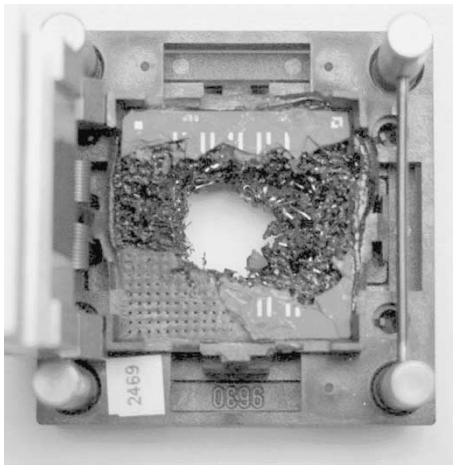

Fig. 1. Test socket can be destroyed by thermal runaway [4].

Stressing during burn-in accelerates the defect mechanisms responsible for early-life failures. Thermal and voltage stresses increase the junction temperature resulting in accelerated aging. Elevated junction temperature, in turn, causes leakages to further increase. In many situations, this may result in positive feedback leading to thermal runaway. Such situations are more likely to occur as technology is scaled to the nanometer regime. Thermal runaway increases the cost of burn-in dramatically. Fig. 1 shows a chip severely damaged by thermal runaway. To avoid thermal runaway, it is crucial to understand and predict the junction temperature under the normal and stress conditions. Junction temperature, in turn, is a function of ambient temperature, package to ambient thermal resistance, package thermal resistance, and static power dissipation. Considering these parameters, one can optimize burn-in environment to minimize the probability of thermal runaway while maintaining the effectiveness of burn-in test.

Section II describes the relevance and types of burn-in. Junction temperature is a critical component in reliability assurance,

Manuscript received October 27, 2003; revised December 23, 2003.

A. Vassighi, O. Semenov, and M. Sachdev are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, N2L 3G1 Canada.

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

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

Digital Object Identifier 10.1109/TDMR.2004.826591

and raising the junction temperature can accelerate several aging mechanisms. Section III describes reliability issues such as gate-oxide breakdown, electromigration and their respective models, and acceleration factors. Sections IV and V focus on the thermal resistance modeling of MOSFETs and junction temperature estimation procedures. Off-state current under a burn-in environment is a critical issue facing the industry. Section VI overviews circuit leakage reduction techniques. Burn-in procedures must be optimized to avoid uncontrolled leakage and subsequent uncontrolled thermal runaway. Section VII describes one such procedure for thermal runaway avoidance. Sections VIII and IX discuss packaging technology and cooling techniques that must evolve to counter the increasing self-heating component in high-performance VLSI circuits.

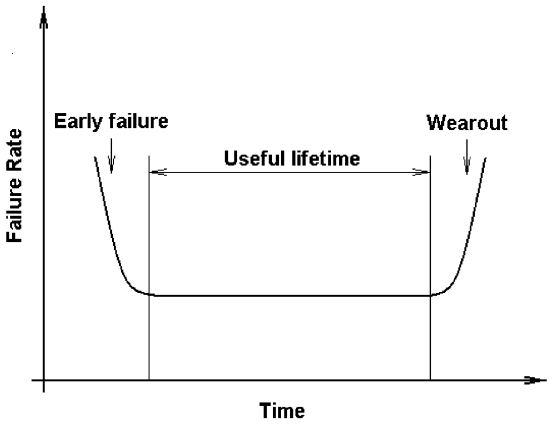

Fig. 2. Bathtub curve.

II. Why Burn-In?

Concurrent technology development and design do not allow technology and design centering to achieve early yield learning and optimal reliability. Weak devices often fail in the field, resulting in early-life failures or "infant mortality." Consequently, IC manufacturers use burn-in procedures to remove weak devices from the population before shipping them to the customer. Fig. 2 shows the bathtub curve indicating the failure rate of electronics devices during their lifetime. Stresses during burn-in cause weak devices to degrade while the ideal normal device population remains unaffected. Standard test programs can detect these degraded devices that will exhibit abnormal voltage or current levels or functional failures.

Analysis of potential failure mechanisms enables the development of good screening processes and tests that are based on the activation of the relevant defect mechanism. Burn-in stress screening is probably the most common technique for detection of infant mortality type of defects due to manufacturing anomalies. This screening typically combines the elevated voltage and elevated temperature to activate the voltage and temperature dependent failure mechanisms for a particular device or process in a relatively short time. Burn-in is used in production of leading edge IC devices to eliminate devices that contain random latent defects, and that have a high probability of early failure in the final application. However, careful attention to the design of stress for burn-in is necessary to ensure that defective devices are stressed to failure, but the useful life of the remaining devices is not adversely affected. The optimization of burn-in stress conditions for constant reliability and reasonable yield loss becomes more difficult for deep-submicron CMOS technologies.

A. Burn-In Procedures

Traditionally, the burn-in procedure is executed prior to a final test procedure that weeds out the parts that have impaired functionality and/or high leakage current from the stresses during burn-in. Burn-in systems are designed to test hundreds of units in parallel over a period of many hours with operating frequency in the tens of megahertz range. There are three basic implementation methods for burn-in: 1) Final package burn-in, where dies are packaged into their final destination packages and are subjected to burn-in at temperatures within the package thermal design constraints; 2) die-level burn-in (DLBI), where dies are placed into temporary carriers before they are actually packaged into their final form, thus reducing the cost of waste associated with added packaging; and 3) wafer-level burn-in (WLBI), where dies are tested while still in wafer form. The last method potentially offers the greatest cost savings by eliminating the packaging waste cost.

The first method offers the most reliable final product since package-related reliability issues are also taken into account. However, this method is expensive since fewer packaged devices can be burned-in simultaneously, and post-burn-in loss includes packaging cost. WLBI is relatively inexpensive, but it results in a relatively less reliable product since packaging-related reliability issues are not addressed. Finally, the die-level burn-in with temporary carriers offers a tradeoff between the other two methods.

B. Static and Dynamic Burn-In

In static burn-in, dies are loaded into burn-in board (BIB) sockets; the BIBs are placed in the burn-in oven. The burn-in system applies power to the devices and heats them to 125 °C–150 °C for periods ranging from 12 to 24 hours. In static burn-in, the device under test (DUT) is powered but not electrically exercised.

Dynamic burn-in mimics the static burn-in process, but also stimulates the DUT's address, data, and clock inputs at a maximum rate (10–30 MHz) determined by the burn-in oven electronics. Under dynamic conditions, circuit nodes are toggled ensuring that voltage stress is applied to various transistors. Neither static nor dynamic burn-in monitors the DUT responses during the stress. Weak dies destroyed by the burn-in process are not detected until a subsequent functional test stage. Recent "intelligent" burn-in systems not only apply power and signals to DUTs; they also monitor DUT outputs.

The test-during-burn-in (TDBI) method can guarantee that devices undergoing burn-in are indeed powered and that input test vectors are being applied. In addition, TDBI can perform some test functions. Detailed information about different burn-in methods and features of burn-in ovens can be found elsewhere [5]–[7].

III. Reliability Issues and Acceleration Factors

The effects of temperature and $V_{\rm DD}$ on microelectronic devices are often assessed by accelerated tests carried out at high temperature and voltage to generate reliability failures in a reasonable time period. Burn-in is often used as a reliability screen to weed out infant mortalities. Weak gate oxides are one of the major components of such failures. These failures are accelerated due to elevated electric field and temperature. Several dielectric breakdown models exist in the literature that can describe intrinsic as well as the defect-related breakdown. In the next section, we consider four widely used models. It is apparent that electric field and junction temperature influence time-to-breakdown of a gate oxide. Metal failures are another typical reliability failure mechanism activated by burn-in. Most metal failures are due to electromigration [8], [9] or stress voiding [9]. In this paper, we consider electromigration as a typical failure mechanism of long time burn-in ($\sim$168 hours) or life testing.

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

The fundamental physical mechanisms of gate-oxide breakdown are divided into two groups: intrinsic and extrinsic oxide breakdown mechanisms. The intrinsic oxide breakdown and wearout refers to defect-free oxide. The failure mechanism can be defined at the critical density of accumulated charge traps in the gate oxide through which a conductive path is formed from one interface to the other. The extrinsic breakdown refers to defects in the oxide whose failure mechanisms are the result of plasma damage, mechanical stress inside of oxide film, contamination, hot carrier damage, or oxide damage by ion implantation. The extrinsic damages in gate oxide typically appear during relatively short-time burn-in testing ($\sim$24 hours). Both breakdown mechanisms appear during burn-in as well as life testing.

The E and $1/E$ models are widely used in intrinsic gate-oxide reliability predictions for oxide thickness $>$50 Å. Both models have a 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, E is the oxide field, E_a is the thermal activation energy, k is Boltzman's constant, and T_j is the junction temperature (K). The E model is based on thermochemical foundation.

On the other hand, 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 electric 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 for calculation and prediction [12]. Hence, starting from about the 180-nm CMOS technology ($T_{\rm ox}$ range is about 26–31 Å), a new TDDB model was proposed [12], [13]. Experiments show that the generation rate of stress-induced leakage current (SILC) and charge to breakdown $(Q_{\rm BD})$ in ultrathin oxides is controlled by gate voltage rather than the electric field. This model (3) includes the gate-oxide thickness $(T_{\rm ox})$ and the gate voltage $(V_{\rm G})$ [14]:

$$T_{\rm BD} = T_0 \cdot \exp\left[\gamma\left(\alpha \cdot T_{\rm ox} + \frac{E_a}{kT_j} - V_{\rm G}\right)\right] \tag{3}$$

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: $(\gamma\cdot\alpha) = 2.0\,1/\text{Å}$, $\gamma = 12.5\,1/\text{V}$, and $(\gamma \cdot E_a) = 575$ meV [14].

Historically, the activation energy has been an independent parameter in gate-oxide breakdown models. However, starting from 130-nm technology, it becomes a function of accelerating electric field, as shown in (4) [15]:

$$E_a \approx 1.15 - 0.07 \cdot E_{\rm ox}\ [\text{eV}]. \tag{4}$$

To explain the TDDB mechanism of extremely thin oxide films ($\sim$20–30 Å), researchers proposed two different approaches: 1) the anode hole injection model [16] and 2) the electron trap generation model [17]. According to the first model, injected electrons generate holes at the anode that can tunnel back into the oxide. Intrinsic breakdown occurs when a critical hole concentration $(Q_{\rm BD})$ is reached. The second model claims that a critical density of electron traps generated during stress is required to trigger oxide breakdown. Based on this model, the breakdown event is presented as the formation of a conductive path of traps connecting the anode to the cathode interface. Recently, it was shown that the anode hole injection model and the electron trap generation model can be directly linked. A new model based on a percolation concept and statistical properties of oxide breakdown was developed [18]. Breakdown can occur only when a connecting path of traps is formed across the gate oxide from the substrate to the gate due to the random defect generation throughout the insulating film. The physics-based analytical model [19], which is the extension and simplification of the common percolation concept, allows us to calculate (5), which is the critical density $(N_{\rm crit})$ of defects per unit of area at breakdown conditions as a function of gate-oxide thickness $(t_{\rm ox})$:

$$N_{\rm crit}^{\rm BD} = \frac{t_{\rm ox}}{\alpha_0^3}\exp\left(-\frac{\alpha_0}{t_{\rm ox}}\ln\left(\frac{A_{\rm ox}}{\alpha_0^2}\right)\right) \tag{5}$$

where α_0 is the lattice constant of a cubic structure in the oxide bulk ($\alpha_0 \approx 2.34$ nm), and A_0 is the oxide area.

The relationship between the charge-to-breakdown $Q_{\rm BD}$, the critical defect density $N_{\rm crit}$, and the injected electron density P_g is [20]

$$Q_{\rm BD} = \frac{qN_{\rm crit}^{\rm BD}}{P_g} \tag{6}$$

The time-to-breakdown of thin oxides is determined by

$$T_{\rm BD} = \frac{Q_{\rm BD}}{J_g} = \frac{qN_{\rm crit}^{\rm BD}}{P_g J_g} \tag{7}$$

where J_g is the tunneling current across the gate oxide. The tunneling current J_g and the injected electron density P_g can be extracted from the experiments using SILC and C–V measurements [20]. Gate-oxide defects have traditionally been a major reason for burn-in. Although other defects are activated during burn-in, it is important to understand the theory of oxide wearout and breakdown.

B. Electromigration (EM)

Interconnect EM is the movement of metal atoms in the direction of electron flow due to momentum transfer from electrons to the metal ions under thermal and voltage stresses. EM is usually modeled by the empirical Black's formula [21], which relates the mean-time-to-failure (MTTF) to the stressing conditions and is given as

$$\text{MTTF} = A \cdot J^{-n} \exp\left(\frac{E_a}{kT_j}\right) \tag{8}$$

where A is the process constant dependent on material and geometry of the metal strip, n is a current exponent factor, T_j is the absolute junction (chip) temperature, k is the Boltzmann's constant, E_a is the activation energy, and J is the current density. The activation energy for Al–Cu metal is in the range of 0.76–0.86 eV [22], and the activation energy for Cu interconnections can vary widely from 0.7–0.9 to 1.0 eV. The lifetime of interconnects is decreased with the reduction of line width [23]. The accuracy of lifetime prediction is strongly dependent on the accuracy of the junction temperature measurement during the acceleration testing.

C. 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 [24]

$$\pi_{\rm T} = 0.1 \exp\left(-A\left(\frac{1}{T_j} - \frac{1}{298}\right)\right) \tag{9}$$

where A is a constant and T_j is the junction temperature (K). Similarly, the voltage acceleration factor is defined in the CNET reliability procedure as [25]

$$\pi_V = A_3 \exp\left[A_4 V_A \left(\frac{T_j}{298}\right)\right] \tag{10}$$

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 (chip) temperature is a fundamental parameter, and should be accurately estimated for each technology generation. To do this, we must understand the properties of new materials and processes used for implementing VLSIs.

IV. Thermal Resistance Models of Semiconductor Devices

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 [26]

$$T_j = T + P \times R_{\rm th} \tag{11}$$

where P is the total power dissipation of the chip and $R_{\rm th}$ is the junction-to-ambient thermal resistance. To estimate the average junction temperature for different technologies, one must investigate the impact of technology scaling on chip power dissipation and thermal resistance. Consequently, we can use (11) to estimate the junction temperature for different technologies.

The initial investigations on technology scaling and thermal resistance were done on bipolar transistors. For these devices, the thermal resistance (°C/mW) was estimated as [27]

$$R_{\rm th} \approx \frac{1}{4K(L \times W)^{\frac{1}{2}}} \tag{12}$$

where K is the thermal conductivity of silicon and $L \times W$ is the emitter size. It was shown that the thermal resistance increased as the emitter size was reduced. Recently, a relationship between thermal resistance of a MOSFET and its geometrical parameters was derived using a three-dimensional (3-D) heat flow equation [28]:

$$R_{\rm th} = \frac{1}{2\pi K}\left[\frac{1}{L}\ln\left(\frac{L+\sqrt{W^2+L^2}}{-L+\sqrt{W^2+L^2}}\right) + \frac{1}{W}\ln\left(\frac{W+\sqrt{W^2+L^2}}{-W+\sqrt{W^2+L^2}}\right)\right] \tag{13}$$

where K is the thermal conductivity of silicon ($K = 1.5 \times 10^{-4}$ W/μm °C [29]), and L and W are channel geometry parameters. This equation was derived for bulk technologies whose substrate thickness was significantly thicker than the thickness of the device layer, and the thermal impedance of the bulk is substantially smaller than that of the device. The thermal conductivity of silicon has a temperature dependence described as [30]

$$K = 154.86 \times \left(\frac{300}{T}\right)^{\frac{4}{3}} \quad (\text{W} \bullet \text{m}^{-1} \bullet \text{K}^{-1}) \tag{14}$$

In our investigations we used $K = 1.5 \times 10^{-4}$ W/μm °C ($T = 300$ K) and assumed that the thermal resistance of silicon is temperature independent [28], [29]. This approximation results in an error of approximately 30% in the solution of linear heat flow differential equations for the temperature range from 25 °C to 125 °C [31], and it is often used in practice [29], [32]. The temperature dependence of silicon thermal conductivity is more important in silicon-on-insulator (SOI) technologies where self-heating contributes to rise in junction temperature. We used the model of (13) for thermal-resistance calculations for MOSFETs in different CMOS technologies.

V. Junction Temperature Estimation

The junction temperature of an IC is defined as the temperature of the silicon substrate, and it is a crucial parameter in

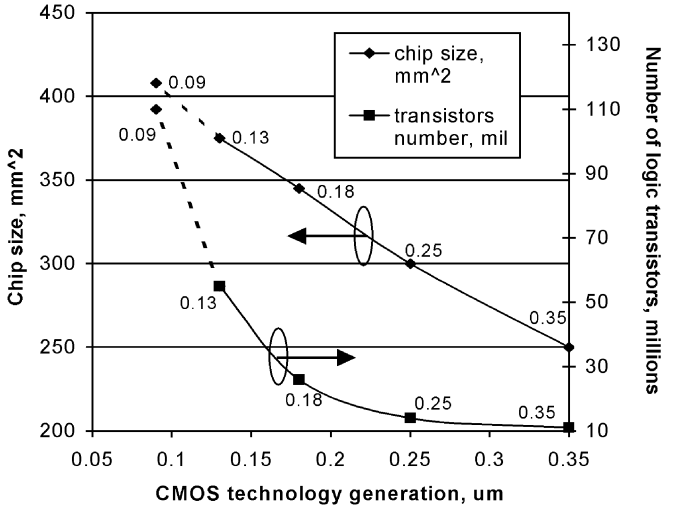

Fig. 3. CMOS technology scaling as reflected in chip size and number of logic transistors [34], [35].

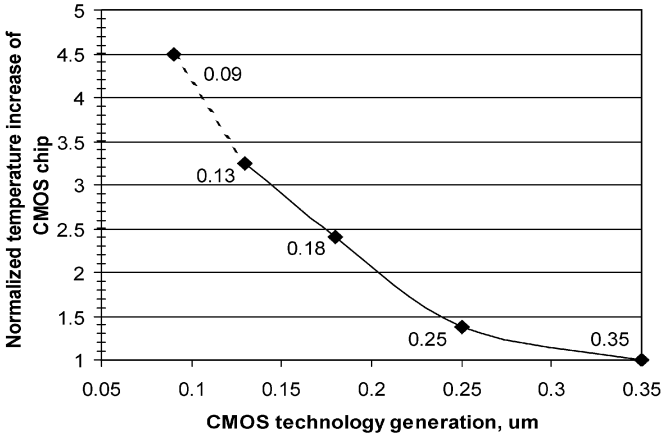

Fig. 4. Normalized chip junction temperature increase with technology [36].

reliability-prediction procedures and burn-in testing. Junction temperature is often a measured value taken from an on-chip sensor. For example, the measured junction temperature of a 1-GHz 64-bit RISC microprocessor implemented in 0.18-μm CMOS technology was reported as 135 °C at $V_{\rm DD} = 1.9$ V [33]. This microprocessor had 15.2 million transistors packed in the 210 mm^2 chip area.

Alternatively, junction temperature can also be estimated from simulations. Equations (11) and (13) can be used for the estimation. Engineers often rely on junction temperature estimated values to develop packaging and cooling solutions for nominal and burn-in conditions due to several reasons. At nominal conditions, the junction temperature prediction will help estimate lifetime warranty for the part and its realistic performance. Similarly, under burn-in conditions, accurate junction temperature estimation may reduce the thermal runaway probability since the margin between optimal burn-in conditions and thermal runaway is reduced as the technology is scaled.

In the subsequent subsections, we show how model-based simulations can be used for junction temperature estimation under nominal and burn-in conditions.

A. Junction Temperature Estimation Under Normal Operating Conditions

Junction temperature increases with technology scaling due to increased transistor density, larger chip size, and increased leakage currents. Fig. 3 shows the increased numbers of transistors and chip size with scaling. A 30% reduction in feature size in each technology scaling results in a doubling of transistor density. At the same time, die sizes are becoming larger. The second curve in Fig. 3 illustrates the die area of high-performance microprocessors. For each successive generation, the increase in area has been between 10% and 20% [34], [35]. These curves allow calculation of the transistor density in a chip for a given technology.

Semenov *et al.* used a four-step procedure to estimate junction temperature in a given technology [36], [37]. They used a 350-nm technology as the reference for normalization. These steps are:

1) estimation of MOSFET power dissipation (P) using simulation and thermal resistance ($R_{\rm th}$) using (13);
2) calculation of the normalized T_j increase over ambient temperature (ΔT) per MOSFET using (11);
3) estimation of MOSFET density (D) from Fig. 3 for the given technology;
4) calculation of normalized T_j increase over ambient temperature for the chip in a given technology as $\Delta T_{\rm chip} = \Delta T \times D$.

The results, shown in Fig. 4, reveal that the normalized temperature increase of the chip is significantly elevated with CMOS technology scaling from 350 to 90 nm under normal operating conditions. The estimated junction temperature of a 90-nm CMOS chip is ~4.5 times higher than the junction temperature of 350-nm CMOS chip. This calculation assumed that the ambient temperature was the same for all analyzed technologies. This nearly exponential increase in chip junction temperature results in an exponential increase in cooling cost [38].

B. Junction Temperature Estimation Under Burn-In Conditions

Burn-in at elevated voltage and temperature conditions is designed to remove the infant mortality IC population from the total population, with little impact on the remaining product population.

In this subsection, we focus on the intrinsic behavior (junction temperature estimation) of the silicon die under burn-in conditions for the sake of simplicity. The thermal impedance network of the package is not considered. For the package level burn-in and DLBI, one must also consider the thermal impedance network of the package [39]. Once this network is known, then (13) can be suitably modified to reflect the total thermal resistance $R_{\rm th}$ of the die and for many types of package. The impact of package thermal resistance on burn-in conditions will be considered in Section VIII.

Estimation of junction temperature under burn-in stress conditions is dependent on several parameters. For example, burn-in yield and reliability are strong functions of the junction temperature. Similarly, package design must also account for the junc-

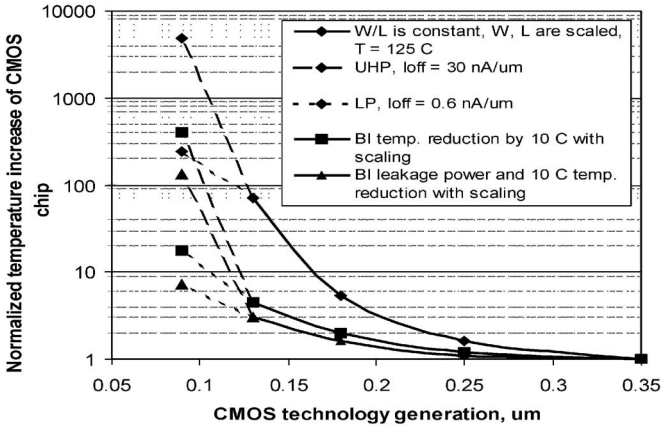

Fig. 5. Normalized chip junction temperature at $VDD + 30\%$ burn-in condition [36].

tion temperature under nominal and stress conditions. Optimization of burn-in conditions such as setting the ambient temperature, stress voltage, cooling considerations, and burn-in duration depend on an accurate estimation of the junction temperature.

The subthreshold leakage becomes the dominant factor in determining the junction temperature in high-performance circuits under stress conditions. Hence, we must estimate the subthreshold leakage current of the chip for the burn-in conditions. The subthreshold conduction of a MOSFET transistor under stress conditions can be explained by a simplified relation illustrated in (15) [40]:

$$I_{\text{subthreshold}} = I_0 e^{\frac{q(V_{\text{G}} - V_{\text{T}})}{kT}} \tag{15}$$

where V_{G} is the gate-to-source voltage, V_{T} is the transistor threshold voltage, q is the electronic charge, k is the Boltzmann's constant, and T is the junction temperature in Kelvin. In the subthreshold region, the term $V_{\text{G}} - V_{\text{T}}$ in (15) is negative. If the junction temperature is linearly increased, it results in an exponential increase in the subthreshold leakage current. Moreover, V_{T} is also a function of the junction temperature and is reduced with increasing temperature resulting in a further increase in the leakage current [40]. Similarly, one can also explain the impact of increased supply voltage on the subthreshold leakage. Elevated drain–source voltage reduces the V_{T} due to the drain-induced barrier lowering (DIBL) effect [41]. As a consequence, the subthreshold leakage is further increased.

The normalized temperature increase of a CMOS chip with scaling at burn-in conditions is shown in Fig. 5 [36]. The curve with the diamond legend depicts the normalized T_j increase if $T = 125\,^{\circ}\text{C}$. For the 90-nm technology, the increase in T_j is different depending on the high-performance or low-power process. If all the transistors are implemented with low V_{T} UHP (ultra high performance) devices (unrealistic), then the normalized T_j is increased by approximately $5000\times$ compared to 0.35-μm CMOS. On the other hand, if all transistors are implemented with LP (low power) devices, then T_j is increased by approximately $230\times$. It should be noted that most of the transistors on chip would be implemented with LP devices.

However, if T is reduced by $10\,^{\circ}\text{C}$ for each technology generation, the normalized T_j is also reduced as shown by the curve with the square legend. Similarly, leakage reduction techniques can further reduce the increased normalized temperature with scaling [42], [43]. Several of these techniques are described in Section VI. It is assumed that the effectiveness of leakage power reduction techniques for high-performance microprocessors should be less than the effectiveness for lower power system-on-chip (SOC) applications mentioned in [43] because these techniques typically reduce the speed of microprocessors. If such techniques are employed as well as reducing T_j by $10\,^{\circ}\text{C}$ for each technology generation, then the normalized T_j increase for 90-nm CMOS with respect to 350-nm CMOS becomes relatively small (7–8×). Despite reduction in T and leakage reduction techniques, the increase in T_j is clearly unacceptable. Obviously, burn-in conditions should be optimized for 130-nm and 90-nm CMOS technologies to reduce the risk of chip over stressing during burn-in.

VI. Leakage Current Reduction Techniques

Major research has been carried out on low power and leakage current reduction [44]. The power consumption in CMOS circuits can be divided into dynamic and static categories. Despite increasing leakage currents with scaling, the dynamic power constitutes the majority of power consumption under normal operational conditions. However, under burn-in conditions, the leakage power becomes significantly large while the operational frequency is reduced drastically. Consequently, the static power component is the dominant part of the total power consumption.

Several circuit techniques have been used to reduce the background leakage current [42]. Some of these techniques can be used during burn-in to restrict the increase in leakage current, and are described below.

Multithreshold Logic: This technique adjusts high-performance critical path transistors with low V_{T} while noncritical paths are implemented with high-V_{T} transistors. Hence, performance and power objectives are achieved at the cost of additional process complexity. Wei *et al.* reported a reduction of more than 80% in leakage power while meeting the performance objectives by using a dual-V_{T} technology [45].

Alternatively, a high-V_{T} transistor can be placed between power supply/ground and the high-performance circuit or block [Fig. 6(a)]. In the active mode, the high-V_{T} transistors are on and since their on-resistance is low, the performance impact is minimal. In the standby mode, the high-V_{T} transistor is off, and hence the leakage is limited to the leakage of a high-V_{T} transistor [46].

Traditionally, multithreshold transistors are realized through different doses of threshold adjust ion implantations. Adjusting the threshold voltages can also be done by depositing two different oxide thicknesses or by different channel lengths [45].

Stack Effect: Another solution to the increasing leakage places a nonstack transistor on a stack of two transistors without affecting the input load [47]. It has been shown that stacking two off-transistors significantly reduces the subthreshold leakage compared to a single off-transistor [Fig. 6(b)]. The drawback of this technique is the increased delay. This delay

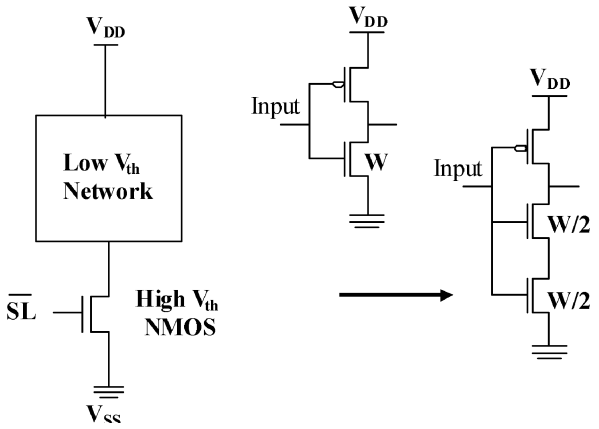

Fig. 6. (a) MTCMOS. (b) Stack effect.

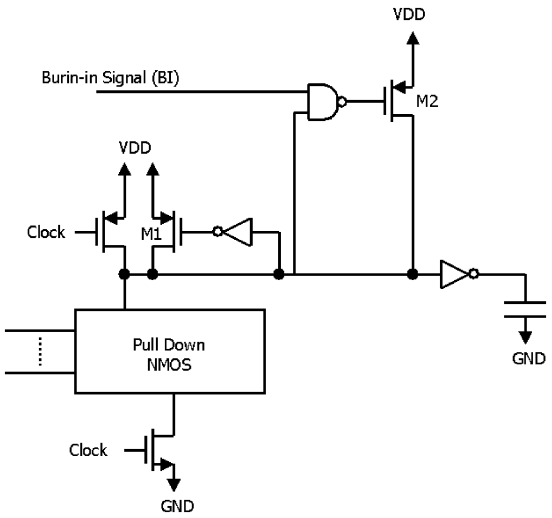

Fig. 7. Burn-in conditional keeper in dynamic circuits [50].

increase is comparable to high-$V_{\rm T}$ logic implementation in a dual-$V_{\rm T}$ technology.

A significantly large fraction of the noncritical path implemented with this technique shows minimal performance degradation while reducing the subthreshold leakage. The stack forcing technique can be either used in conjunction with dual-$V_{\rm T}$ or with a single-$V_{\rm T}$ technology [47].

Reverse Body Bias (RBB): This is another technique to reduce leakage current during active operation, burn-in, and standby mode. During active operation, RBB is applied to the idle portion of the chip to reduce overall chip leakage power without impacting the performance. Since in the chip, operational frequency is very low during burn-in, RBB can be applied to the whole chip simultaneously.

Although increasing RBB reduces the weak inversion current monotonically, the junction leakage component increases with larger RBB due to the gate-induced drain leakage (GIDL) effect. An optimal point is achieved where any further increase in RBB does not produce an overall subthreshold current reduction. The effectiveness of RBB diminishes with scaling. Keshavarzi *et al.* showed that the maximum leakage reduction through RBB is diminished from 4–5$\times$ in 180-nm technology to 2–3$\times$ in 130-nm technology [48]. However, the effectiveness of RBB can be improved by using a dual-$V_{\rm DD}$ design, as has been reported in [49].

Conditional Keepers: Degradation of dynamic circuit functionality is a problem during burn-in testing because of high leakages in stress conditions. To overcome this problem, a keeper technique was proposed that is active during the burn-in, and is inactive during normal operating mode. Consequently, the dynamic circuit remains functional under burn-in without relaxing the maximum burn-in stress and without any significant performance degradation under normal operating conditions [50].

The elevated temperature and voltage exponentially increase the leakage current. The large leakage current can discharge dynamic nodes resulting in incorrect operation of dynamic circuits. Conditional burn-in keepers are designed for functionality of sub-130-nm dynamic circuits. The conditional keeper technique uses an extra keeper for the burn-in mode to compensate for higher leakage in burn-in. Fig. 7 shows this technique. Transistor M1 is the standard keeper, while transistor M2 is the burn-in keeper. M2 is off in the normal operating condition and turns on for the burn-in mode using a burn-in signal through the NAND gate.

VII. Thermal Runaway Avoidance

Several reliability failure mechanisms are accelerated by elevated temperature. These mechanisms include metal stress voiding and electromigration, metal sliver bridging shorts, contamination, and gate-oxide wear out and breakdown [51]. However, there are physical and burn-in equipment-related limitations for junction temperature and voltage stress. Die failure rate (failures per million) increases exponentially with junction temperature for most failure mechanisms [52]. As a result, the yield loss may increase if the burn-in conditions cause overstress. In a limiting case, an unabated increase in the junction temperature may lead to the thermal runaway. Hence, the junction temperature should be optimized for normal and burn-in conditions.

A. Physical and Practical Limits of Junction Temperature

The maximum operating temperatures for semiconductor devices can be estimated from semiconductor intrinsic carrier density, which depends on the bandgap of the material. When the intrinsic carrier density reaches the doping level of the active region of devices, then the electrical parameters change drastically. The highest operating junction temperature for standard silicon technology is about 200 °C, however the circuit performance is reduced substantially [53]. The influence of temperature on some important MOSFET parameters is summarized in Table I.

Several practical considerations limit the junction temperature to a much lower value. A limit of 150 °C for junction temperature is often used for VLSI ICs [34]. The peak junction temperature of a PowerPC microprocessor implemented in a 0.35-μm CMOS technology was reported to be approximately 90 °C–100 °C at an operating speed of 200–250 MHz [54], [55].

B. A Procedure for Thermal Runaway Avoidance

The increase of IC background leakage current due to the CMOS technology scaling, especially under burn-in conditions,

TABLE I
TEMPERATURE DEPENDENCE OF IMPORTANT Si-MOSFET PARAMETERS, DATA ADOPTED FROM [53]

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

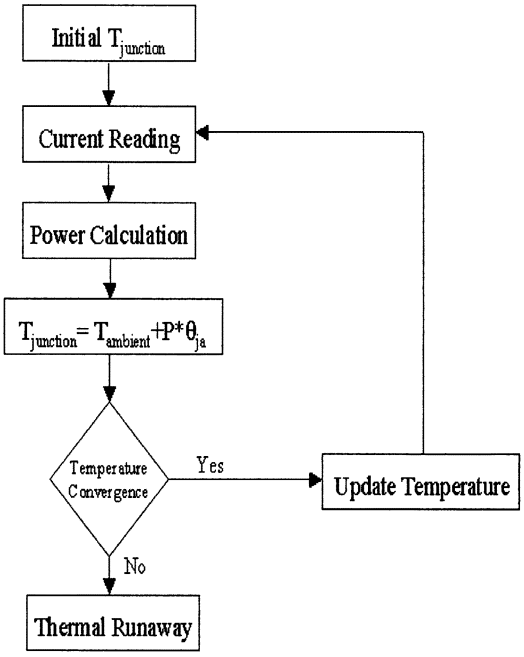

Fig. 8. A procedure for junction temperature estimation [56]. **<PLEASE PROVIDE A CITATION FOR FIG. 8 IN THE TEXT.>**

leads to a requirement to assess the ambient versus safe junction temperature conditions, since they relate to IC thermal runaway.

Vassighi *et al.* described one such procedure whose program flow chart is shown in [56]. For the initial junction temperature, the input current for a single transistor is the input to the program. Based on the circuit implementation and architecture, the total power is computed and junction temperature is updated in (11). Using this procedure for any given voltage and process technology, junction temperature is calculated and convergence of the obtained temperature is tested. After several iterations, the junction temperature will either converge to a stable value, or it will increase and lead to chip thermal runaway.

A 32-bit microprocessor in 100-nm dual-V_T CMOS technology was used to verify the procedure. The parameters of this program were calibrated to the experimental data from the microprocessor. Fig. 9 shows the electrothermal simulation results carried out for a 32-bit microprocessor implemented in a 100-nm technology. The dashed and solid graphs represent junction temperatures in air-cooled and liquid-cooled burn-in ovens,

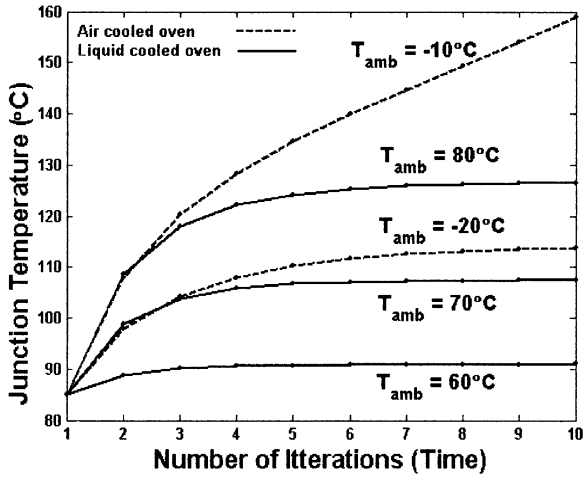

Fig. 9. Junction temperature in air-cooled versus liquid-cooled burn-in ovens under given ambient conditions [56].

respectively, under given ambient temperature conditions. The power supply voltage was set to $1.35 \times V_{DD}$.

Fig. 9 clearly shows that the ambient temperature in air-cooled ovens must be kept at $-20\,^{\circ}\mathrm{C}$ or lower to stabilize the junction temperature and hence avoid thermal runaway. However, it is unrealistic to achieve ambient temperature lower than room temperature in air-cooled burn-in ovens. On the other hand, liquid-cooled ovens are more efficient and allow a higher ambient temperature due to their relatively lower junction to ambient thermal resistances. Hence, they are better suited to avoid the thermal runaway under burn-in conditions.

VIII. IMPACT OF PACKAGE THERMAL RESISTANCE ON BURN-IN

High-performance VLSI circuits such as microprocessors significantly challenge power delivery and heat removal due to smaller dimensions and increasing power dissipation. Technical challenges in the thermal management of microprocessors arise from two causes [57]: 1) increased dynamic and leakage power dissipation associated with technology scaling and 2) heat removal from localized hot spots. The former is especially important for burn-in since the leakage power is exponentially increased under stress conditions. Typically, thermal management features are integrated in packages to spread heat from die to the heat sink. The heat sink dissipates the heat into local environments.

A typical thermal resistance network of a packaged die is shown in Fig. 10. By definition, the case temperature (T_c) is the temperature at the external surface of the package. All semiconductor packages have multiple elements. In the simplest form these elements include the semiconductor die, thermal interface material, and the heat sink base. The thermal conductivity of these package elements for the Pentium III Xeon microprocessor is given in Table II.

In a common case, the junction temperature increase over ambient temperature has three components [59]:

$$\Delta T = P \times [R_{\mathrm{th}}(\text{die-pack}) + R_{\mathrm{th}}(\text{pack-sink}) + R_{\mathrm{th}}(\text{sink-amb})] \quad (16)$$

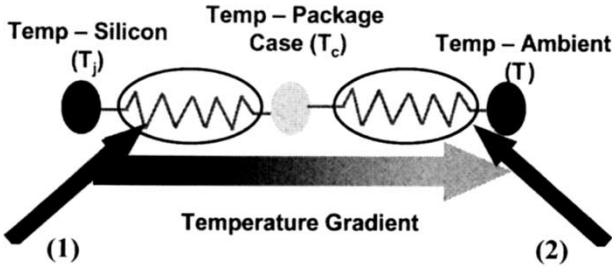

Fig. 10. Thermal resistance network of a packaged die. (1) Junction to case (package). (2) Case to ambient (heat sink) [57].

TABLE II
THERMAL CONDUCTIVITY OF PACKAGE COMPONENTS [58]

Package component	Conductivity, W/mK
Silicon die	120
Thermal interface material	3.8
Heat sink base	180

where $R_{\rm th}$(die-pack), $R_{\rm th}$(pack-sink), and $R_{\rm th}$(sink-amb) are die to package, package to heat sink, and heat sink to ambient thermal resistances, respectively, and P is the total power dissipation of the chip. The first component in (16) was considered in previous sections. The third component is determined by the cooling techniques and will be considered in the next section. Here, we consider the second component in (16), which can be rewritten as:

$$\Delta T(\text{Package}) = P_{\rm MOSFET} \cdot \frac{D}{2} \cdot R_{\rm th}(\text{pack-sink}) \qquad (17)$$

where $P_{\rm MOSFET}$ is the transistor power dissipation and D is the transistor density. The package to heat sink thermal resistance, $R_{\rm th}$(pack-sink), is crucial in removing heat during burn-in. Values of 0.9–1.2 °C/W were reported for $R_{\rm th\text{-}PH}$ in 350-nm technology [60], [61]. It is predicted that a reduction of approximately 22% in $R_{\rm th}$(pack-sink) per technology generation is required to just compensate the increased power density with technology scaling [62]. Fig. 11 shows these projections for the 350-nm technology to 90-nm technology.

IX. COOLING TECHNIQUES FOR BURN-IN

Low-power devices can be burned-in without attention to thermal considerations. However, as power dissipation increases with technology scaling for high-performance chips, burn-in requires advanced cooling concepts and additional hardware to facilitate direct contact between the heat sink and the die. Advanced burn-in ovens should provide uniform temperature distribution in the chamber and precise temperature control for each individual device. The power dissipation within one lot of devices can vary by ±40% due to manufacturing variations and different test vectors applied during burn-in. This variation in power, and approximately ±30% variation in oven airflow, can create a significant variation in package temperature [63]. If the device becomes too hot, it may be damaged while other devices may not be adequately burned-in. To uniformly stress all devices, each package device temperature must be kept close to the specified burn-in temperature. This

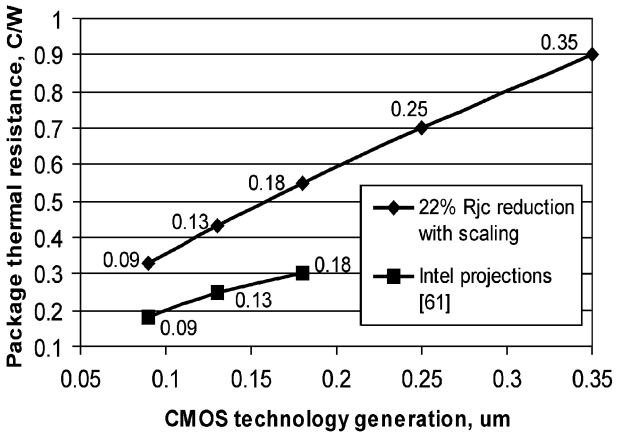

Fig. 11. Reduction of package thermal resistance with technology scaling.

is achieved by developing advanced cooling techniques and burn-in boards with embedded thermal sensors.

A. 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 oven may contain several hundred dies. 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 [7]. We can use the power dissipation of a single transistor in an inverter at static stressed conditions and the number of transistors of the logic chip to estimate different CMOS technologies. We can then estimate the maximum number of die for different technologies that can be simultaneously powered in a burn-in oven using (18).

$$N_{\rm dies} = \frac{P_{\rm oven}}{P_{\rm transistor} \times \frac{N_{\rm transistors}}{2}} \qquad (18)$$

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 (18) assumes that 50% of the total number of transistors are off at any point during burn-in assuming fully static CMOS design. Results are shown in Fig. 12.

Burn-in ovens, such as the PBC1-80 of Despatch Industries [7] and Max-4 of Aehr Test Systems [5], have maximum power dissipation of about 2500 and 15 000 W, respectively, at 125 °C. The room ambient temperature is assumed to be 25 °C.

B. Air-Cooling Technique

For CMOS IC technologies of 0.35 μm and above, generally IC junction heating during burn-in has not been a major issue, and the oven temperature could be set to eliminate the temperature-related overstress. However, for 0.25-μm technology and below, device self-heating has been described to become a more significant issue, and air-cooling techniques began to be implemented to remove heat from each device and the oven.

Air-cooled burn-in ovens are reasonably effective in heat removal from devices dissipating up to 30–40 W [63]. Often, an

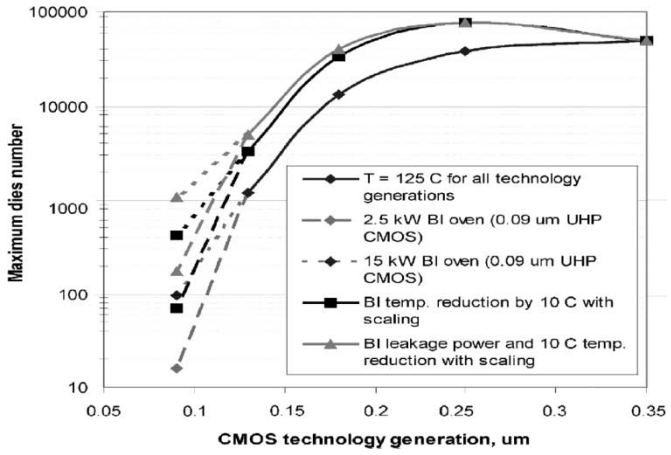

Fig. 12. Maximum dies number for one burn-in load versus CMOS technology scaling.

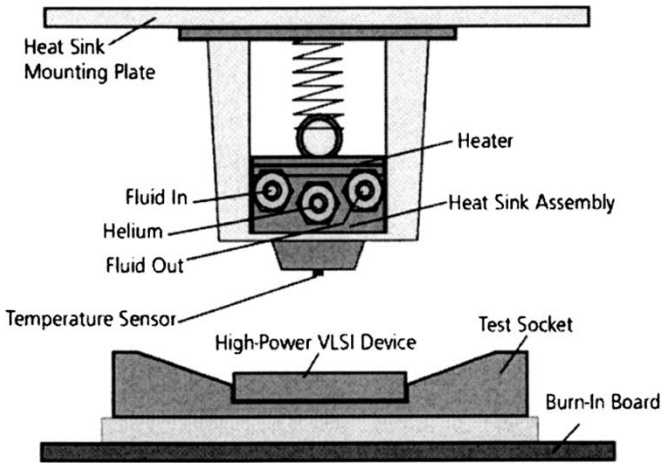

Fig. 13. Water-cooled heat sink, adopted from [63].

air-cooled heat sink and embedded thermal sensors are used to control the individual temperature of device. The air temperature and air velocity are dependent on the device power, the overall thermal resistance of the heat sink assembly and burn-in socket, and the required package temperature. The air temperature and velocity must be controlled so that the embedded heat sink can limit the device temperature increase over the range of heat dissipation. The device temperature can be controlled in the range of 50 °C–150 °C with an accuracy of ± 3 °C [63]. Device temperature is usually measured by attaching a small thermocouple directly on the device or by using sensors integrated into the device [64].

Another air-cooling technique was developed for device power dissipation from 35 to 75 W [63]. This approach uses a small fan mounted above the heat sink of each device. The amount of allowable device power dissipation is a function of the air temperature, air velocity, thermal resistances of the heat sink, and the package.

To ensure quality output, ovens are designed to ensure that the temperature distribution across all the boards is uniform and adequate. The level and uniformity of the temperature across the burn-in boards is controlled by the total airflow induced in the oven and the uniformity of the airflow distribution between the boards. The design of an airflow network becomes increasingly more complicated as device power dissipation increases [65].

C. Liquid-Cooling Technique

As power dissipation increases beyond 75 W per device, the thermal resistance of package to ambient must be lowered to allow removal of excess heat. Air-cooling burn-in techniques are not effective for power dissipation in this range, and this has fostered the development of liquid-based cooling techniques. Fig. 13 illustrates one such technique [63]. A temperature sensor embedded in the heat sink measures the device temperature. Helium is injected into the heat sink to provide a lower thermal interface between the device and the heat sink. This technique lowers the heat sink to ambient thermal resistance by approximately 40%.

Each heat sink has a temperature-controlled heater. The burn-in ovens with liquid-cooled heat sinks can burn-in devices that dissipate over 150 W of power [66], [67]. In such ovens, the ambient temperature for each device can be optimized for optimal burn-in conditions. This is important since self-heating dissipation can vary significantly due to inherent process spreads in scaled technologies. The thermal control during test and burn-in of devices with high leakage power dissipation (above 75 W) plays a key role in increasing the post-burn-in yield.

Special thermal test chips and modules were developed to measure temperature gradients in packages and heat sinks in burn-in equipment [65], [66]. For example, IBM used a TV994 thermal test chip for burn-in equipment qualification. This 14.7-mm^2 chip has nine small resistive temperature detectors (RTD) and four large heater resistors, one covering each quadrant of the chip [66]. The thermal interface tests evaluate temperature gradients within the device and between the device and heat sink. Temperature differences are normalized with respect to applied device power. The test is used to optimize and evaluate factors such as heat sink material, flatness and various properties of interface pads, and liquids and gases that can be between the chip and heat sink.

X. Burn-In Optimization for Yield and Reliability

Yield and reliability are two important factors in semiconductor manufacturing. Typically, three parameters significantly affect yield and reliability of ICs [68]: 1) design-related parameters (chip area and gate-oxide thickness); 2) process-related parameters (defect distribution and density); and 3) operation-related parameters (voltage and temperature). It has been experimentally verified that defects that cause burn-in failures (early-life reliability failures) are fundamentally the same in nature as defects that cause wafer probe failures (yield failures) [69], [70].

Researchers have also identified two key reliability indicators in order to optimize yield during burn-in: 1) local region yield and 2) the number of defects that have been repaired (for chips containing redundancy). Experimentally, it has been shown that die with many faulty neighbors can pose a significantly greater early-reliability risk than chips with few faulty neighbors [71].

An IC with a redundancy-related repair is more likely to have a latent defect mechanism resulting in early-life failure [69].

The key to optimizing burn-in lies in identifying those die that are most likely to fail during burn-in before the burn-in is actually performed. Once identified, dies of higher reliability risk may be subjected to more rigorous testing (longer burn-in duration), while those dies deemed more reliable may have a reduced stress, or no stress at all. Barnett *et al.* proposed the post-burn-in yield model, which includes the burn-in time as a parameter [71]. It was assumed that the average number of latent defects (λ_{L}) per chip is time dependent as follows:

$$\lambda_L(t) = \alpha \times \gamma \times \left(1 - Y_K^{\frac{1}{\alpha}}\right) \times \left(\frac{t}{\tau}\right)^{\beta} \tag{19}$$

where α is the defect clustering parameter, $\gamma \approx 0.01 - 0.02$ is the fitting parameter, Y_K is the wafer test yield (yield before burn-in), τ is the burn-in time in hours, and β is the shape parameter of Weibull distribution of the reliability function. The post-burn-in reliability yield (i.e., the number of dies surviving burn-in) is modeled as follows:

$$R(t) = \left[1 + \frac{\lambda_L(t)}{\alpha}\right]^{-\alpha}. \tag{20}$$

Kim *et al.* [72] developed another model for post-burn-in reliability R and yield loss Y_{loss} as shown in the equations below. They assumed that the gate-oxide damage is the leading defect mechanism.

$$Y_{\mathrm{loss}} = Y\left(1 - Y^{\frac{v}{1-v}}\right) \tag{21}$$

$$R = Y^{\frac{1}{(1-u)^2-1}} \tag{22}$$

where Y is the yield before burn-in, v is a constant dependent on the burn-in time, stress voltage, and temperature, and is related to the gate-oxide damage, incurred during burn-in. On the other hand, u is a constant dependent on operating voltage and time, and is related to the gate-oxide damage incurred during normal device operation. The typical range of v and u constants is from 0.1 to 0.9. Vassighi *et al.* used the $1/E$ gate-oxide breakdown model and the above-mentioned post-burn-in yield loss model to demonstrate that the post-burn-in yield loss increases exponentially with elevation of stress temperature for a given stress voltage [73]. This result was obtained for a 180-nm CMOS technology ($T_{\mathrm{ox}} = 41$ Å).

Burn-in removes the infant mortality, hence improving the outgoing device reliability. However, burn-in may affect the post-burn-in yield of ICs since latent defects may become enhanced during burn-in, with a resultant increase in post-burn-in yield loss. The amount of yield loss depends on burn-in conditions (voltage, temperature, time). Since the stress voltage and the stress temperature provide the acceleration during burn-in, the burn-in time is the parameter that is manipulated to control the post-burn-in yield loss using above mentioned models. In practice, many IC manufactures reduce the burn-in time to 10 hours or even skip burn-in, when the yield before burn-in is high (∼98%) and burn-in escapes are low (∼100 PPM) [72]. The amount of burn-in escape is estimated by the early failure rate test, which is performed on 10 000 final products from at least three lots with duration of approximately 12–48 hours under burn-in conditions.

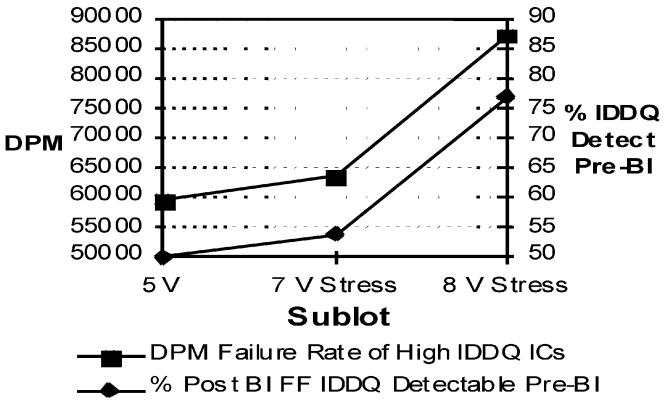

Fig. 14. I_{DDQ} detection of burn-in functional failures and defect level of ICs that failed only I_{DDQ} tests [79].

XI. Burn-In Elimination

The elimination of burn-in by an alternate screening method has been a long sought after goal. Despite the expense, mechanical and EOS/ESD damage to the burn-in parts, and lengthened time to market, burned-in parts typically achieve a better quality measure than nonburned-in parts. These negative features of burn-in stimulated a search for screening methods that might achieve the same lowering of DPM (defects-per-million) levels of shipped parts. In the pre-nanometer technologies where transistor channel lengths were above the 0.35 μm level, the I_{DDQ} test was reported by several companies as successful in eliminating or reducing burn-in [74]–[78]. Intel reported experiments on several thousand ICs and found that I_{DDQ} when combined with a short high voltage stress on the parts, yielded near zero DPM outgoing quality levels [74]. Kawasaki Steel reported a similar study using several hundreds of thousands of parts showing that I_{DDQ} screens could eliminate burn-in [75]. LSI Logic and Philips Semiconductors reported similar success with I_{DDQ} screening to eliminate burn-in [76], [77]. McEuen of Ford Microelectronics reported that nominal voltage I_{DDQ} testing enabled reduction of burn-in failures by 51% [78].

However, one caveat of these reports was that I_{DDQ} screening was successful in burn-in elimination only if the manufacturing quality levels were high. I_{DDQ} could not eliminate burn-in on rogue lots. This obstacle was overcome in a study funded jointly by Sandia National Labs and the Sematech organization [79]. The experiment used 3495 parts in a dynamic burn-in that separated the parts into a control sample, a 7-V stress sample, and an 8-V stress sample. 40 000 I_{DDQ} measurements were taken per die during the control and voltage stress sample tests. I_{DDQ} test limits were set tight at the $\pm 3\sigma$ levels from the mean plus a tester noise guard band. Fig. 14 summarizes the prediction of functional failure during burn-in from pre-burn-in I_{DDQ} test data. The I_{DDQ} screen predicted that I_{DDQ} testing would detect 50% of the control parts (5 V), 54% of the 7-V stressed parts, and 77% of the 8-V stressed parts. DPM of the data showed that the DPM level of the control group was 1.75 times larger than the 8-V stressed

sample. Cost models also showed economic justification of the I_{DDQ} test in eliminating burn-in.

A test methods study was also funded by Sematech with IBM, and that was the only study to date that stated that I_{DDQ} testing did not show elimination of burn-in [80]. However, no explanation was given as to why the data contradicted the several reports that it would, and no burn-in data were given.

While these experiments demonstrated that parametric measurements could be used to eliminate burn-in, they were done on long channel transistor ICs whose background noise levels obscured sensitive I_{DDQ} or other parametric measurements. The question is, how does I_{DDQ} or other parametric measurements perform for nanometer CMOS ICs. There are two public reported of success. The first was at a burn-in panel at the International Reliability Physics Symposium (IRPS) in 2001 [81]. Panelists from five major companies said that if the manufacturing quality of the lots could be measured as high, then parametric screens could achieve burn-in elimination. They stressed that this approach did not work if the quality levels were not high.

The second report on nanotechnology parts came from a team from LSI Logic and Portland State University [82]–[84]. They reported parametric screening of outlier parts using post-test statistical processing methods on the whole wafer data. The technique measures statistics of neighboring or other die locations on the wafer to determine I_{DDQ} and $V_{\rm DD}{\rm Min}$ (lowest functional voltage $V_{\rm DD}$) test limits. This study reported the application of post-test statistics to burn-in elimination, but did not specifically report burn-in elimination data. The severe problems that nanometer ICs present to burn-in make these parametric screening techniques of high interest.

XII. Conclusion

Burn-in is a quality improvement procedure widely used for high-performance and high-volume products. This paper provides an overview of CMOS technology scaling and its impact on burn-in.

Smaller geometries, increased transistor leakages, and larger integrations are resulting in higher junction temperatures and self-heating. Elevated junction temperature, in turn, causes leakages to increase further. In many situations, this may result in positive feedback leading to thermal runaway. Therefore, burn-in leakage reduction techniques, thermal runaway avoidance procedures must be evolved. Moreover, deep-submicron devices will require advance packaging and liquid cooling techniques to lower the junction to ambient thermal resistance.

In scaled technologies, burn-in optimization for yield and reliability will be of crucial significance owing to larger number of design and technology variables. In some situations, individual chip level burn-in optimization will be necessary in order to provide optimum burn-in environment for each chip. Significant research has been carried out toward burn-in elimination. For long channel devices, several companies have reported burn-in elimination with I_{DDQ} under controlled process conditions. However, it appears to be difficult for deep-submicron technologies.

References

[1] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, pp. 23–29, July–Aug. 1999.

[2] S. Rusu, "Trends and challenges in VLSI technology scaling toward 100 nm," presented at the ESSCIRC 2001, http://www.esscirc.org/esscirc2001/C01_Presentations/404.pdf.

[3] S. Thompson, P. Packan, and M. Bohr, "MOS scaling: transistor challenges for the 21st century," *Intel Tech. J.*, vol. Q3, pp. 1–19, 1998. http://developer.intel.com/technology/itj/archive.htm.

[4] M. Miller, "Next generation burn-in and test systems for Athlon microprocessors: hybrid burn-in," in *AMD Burn-In and Test Socket Workshop, Session 5*, 2001.

[5] Aehr Test Systems [Online]. Available: http://www.aehr.com

[6] Micro Control Co. [Online]. Available: http://www.microcontrol.com/

[7] Despatch Industries [Online]. Available: http://www.despatch.com/pdfs/PBC.pdf

[8] E. T. Ogawa, K.-D.Ki-Don Lee, V. A. Blaschke, and P. S. Ho, "Electromigration reliability issues in dual-damascene Cu interconnections," *IEEE Trans. Reliabil.*, vol. 51, pp. 403–419, Dec. 2002.

[9] C. F. Hawkins, A. Keshavarzi, and J. M. Soden, "Reliability, test and $Iddq$measurements," in *IEEE Int. Workshop on I_{DDQ} Testing*, 1997, pp. 96–102.

[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," in *Proc. IEEE Int. Reliability Physics Symp.*, 2000, pp. 7–15.

[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 ($T_{\rm ox} < 32$ Å)," in *Proc. IEEE Int. Reliability Workshop*, 2001, pp. 20–25.

[15] M. Kimura, "Field and temperature acceleration models for time-dependent dielectric breakdown," *IEEE Trans. Electron Devices*, vol. 46, pp. 220–229, Jan. 1999.

[16] I. C. Chen, S. Holland, K. K. Young, C. Chang, and C. Hu, "Substrate hole current and oxide breakdown," *Appl. Phys. Lett.*, vol. 49, no. 11, pp. 669–671, 1986.

[17] P. P. Apte and K. C. Saraswat, "Modeling ultra thin dielectric breakdown on correlation of charge trap-generation to charge-to-breakdown," in *Proc. IRPS*, 1994, pp. 136–142.

[18] R. Degraeve, G. Groeseneken, R. Bellens, J. L. Ogier, M. Depas, P. J. Roussel, and H. E. Maes, "New insights in the relation between electron trap generation and the statistical properties of oxide breakdown," *IEEE Trans. Electron Devices*, vol. 45, pp. 904–911, Apr. 1998.

[19] J. Sune, "New physics-based analytical approach to the thin-oxide breakdown statistics," *IEEE Electron Device Lett.*, vol. 22, pp. 296–298, June 2001.

[20] J. H. Stathis, "Physical and predictive models of ultra thin oxide reliability in CMOS devices and circuits," in *Proc. IRPS*, 2001, pp. 132–149.

[21] J. Black, "Electromigration—a brief survey and some recent results," *IEEE Trans. Electron Devices*, vol. ED-16, pp. 338–347, Apr. 1969.

[22] W. B. Loh, M. S. Tse, L. Chan, and K. F. Lo, "Wafer-level electromigration reliability test for deep submicron interconnect metallization," in *Proc. IEEE Hong Kong Electron Device Meeting*, 1998, pp. 157–160.

[23] C.-K. Hu, R. Rosengerg, H. S. Rathore, D. B. Nguyen, and B. Agarwala, "Scaling effect on electromigration in on-chip Cu wiring," in *Proc. IEEE Int. Interconnect Technology Conf.*, 1999, pp. 267–269.

[24] P. Lall, "Tutorial: temperature as an input to microelectronics-reliability models," *IEEE Trans. Reliabil.*, vol. 45, pp. 3–9, Jan. 1996.

[25] J. B. Bowles, "A survey of reliability-prediction procedures for microelectronics devices," *IEEE Trans. Reliabil.*, vol. 41, pp. 2–12, Jan. 1992.

[26] P. Tadayon, "Thermal challenges during microprocessor testing," *Intel Technol. J.*, vol. Q3, pp. 1–8, 2000.

[27] 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.

[28] N. Rinaldi, "Thermal analysis of solid-state devices and circuits: an analytical approach," *Solid-State Electron.*, vol. 44, no. 10, pp. 1789–1798, 2000.

[29] ——, "On the modeling of the transient thermal behavior of semiconductor devices," *IEEE Trans. Electron Devices*, vol. 48, pp. 2796–2802, Dec. 2001.

[30] D. L. Blackburn and A. R. Hefner, "Thermal components models for electro-thermal network simulation," in *Proc. 9th IEEE SEMI-THERM Symp.*, 1993, pp. 88–98.

[31] G. Digele, S. Lindenkreuz, and E. Kasper, "Fully coupled dynamic eletro-thermal simulation," *IEEE Trans. VLSI*, vol. 5, pp. 250–257, Mar. 1997.

[32] S. Im and K. Banerjee, "Full chip thermal analysis of planar (2D) and vertically integrated (3D) high-performance ICs," in *Proc. IEDM*, 2000, pp. 727–730.

[33] J. Ahn, H.-S. Kim, T.-J. Kim, H.-H. Shin, Y.-H.Y.-Ho 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.

[34] International Technology Roadmap for Semiconductors (ITRS). [Online]. Available: http://public.itrs.net/

[35] D. P. Vallett and J. M. Soden, "Finding fault with deep-submicron ICs," *IEEE Spectrum*, vol. 34, pp. 39–50, Oct. 1997.

[36] 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.*, 2003, pp. 95–104.

[37] ——, "Effect of CMOS technology scaling on thermal management during burn-in," *IEEE Trans. Semicond. Manufact.*, vol. 16, pp. 686–695, Apr. 2003.

[38] S. H. Gunter, F. Binns, D. M. Carmean, and J. C. Hall, "Managing the impact of increasing microprocessor power consumption," *Intel Technol. J.*, vol. Q1, pp. 1–9, 2001. http://developer.intel.com/technology/itj/archive.htm.

[39] G. Kromann, "Thermal management of a C4/CBGA interconnect technology for a high-performance RISC microprocessor: the Motorola PowerPC 620 microprocessor," in *Proc. IEEE Electronic and Technology Conf.*, 1996, pp. 652–659.

[40] S. M. Sze, *Semiconductor Device Physics and Technology*, 2nd ed. New York: Wiley, 2002.

[41] A. N. Mutlu and M. Rahman, "Two-dimensional analytical model for drain induced barrier lowering (DIBL) in short channel MOSFETs," in *Proc. Southeastcon 2000 Conf.*, 2000, pp. 340–344.

[42] S. Borkar, "Leakage reduction in digital CMOS circuits," in *Proc. IEEE Solid-State Circuits Conf.*, 2002, pp. 577–580.

[43] A. B. Kahng, "ITRS-2001 design ITWG," in *Proc. ITRS Release Conf.*, 2001. http://public.itrs.net/Files/2001WinterMeeting/Presentations/Design.pdf.

[44] K. Roy, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. IEEE*, vol. 91, pp. 305–327, Feb. 2003.

[45] L. Wei, Z. Chen, K. Roy, M. C. Johnson, Y. Ye, and V. K. De, "Design and optimization of dual-threshold circuits for low-voltage low-power applications," *IEEE Trans. VLSI Syst.*, vol. 7, pp. 16–24, Jan. 1999.

[46] L. Wei, K. Roy, and V. K. De, "Low voltage low power CMOS design techniques for deep submicron ICs," in *Proc. Int.. Conf. VLSI Design*, 2000, pp. 24–29.

[47] S. Narendra, S. Borkar, V. De, D. Antoniadis, and A. Chandrakasan, "Scaling of stack effect and its application for leakage reduction," in *Proc. Int. Symp. Low Power Electronics and Design (ISLPED)*, 2001, pp. 195–200.

[48] A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De, "Effectiveness of reverse body bias for leakage control in scaled dual Vt CMOS ICs," in *Proc. Int. Symp. Low Power Electronics and Design (ISLPED)*, 2001, pp. 207–212.

[49] L. T. Clark, N. Deutscher, S. Demmons, and F. Ricci, "Standby power management for a 0.18 um microprocessor," in *Proc. ISLPED*, 2002, pp. 7–12.

[50] A. Alvandpour, R. Krishnamurthy, S. Borkar, A. Rahman, and C. Webb, "A burn-in tolerant dynamic circuit technique," in *Proc. IEEE Custom Integrated Circuits Conf.*, 2002, pp. 81–84.

[51] 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.

[52] N. F. Dean and A. Gupta, "Characterization of a thermal interface material for burn-in application," in *Proc. IEEE Thermal and Thermomechanical Phenomena in Electronic Systems*, 2000, pp. 36–41.

[53] W. Wondrak, "Physical limits and lifetime limitations of semiconductor devices at high temperature," *Microelectron. Reliabil.*, vol. 39, no. 6–7, pp. 1113–1120, 1999.

[54] H. Sanchez, B. Kuttanna, T. Olson, M. Alexander, G. Gerosa, R. Philip, and J. Alvarez, "Thermal management system for high performance PowerPC microprocessors," in *Proc. IEEE COMPCON*, 1997, pp. 325–330.

[55] G. Gerosa, M. Alexander, J. Alvarez, C. Croxton, M. D'Addeo, A. R. Kennedy, C. Nicoletta, J. P. Nissen, R. Philip, P. Reed, H. Sanchez, S. A. Taylor, and B. Burgess, "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.

[56] A. Vassighi, O. Semenov, M. Sachdev, and A. Keshavarzi, "Thermal management of high performance microprocessors in burn-in environment," in *Proc. 18th IEEE Int. Symp. Defect and Fault Tolerance in VLSI Systems*, 2003.

[57] R. Mahajan, R. Nair, V. Wakharkan, J. Swan, J. Tang, and G. Vandentop, "Emerging directions for packaging technologies," *Intel Technol. J.*, vol. 6, no. 2, pp. 62–75, 2002. http://developer.intel.com/technology/itj/2002/volume06issue02/.

[58] T. J. Goh, K. N. Seetharamu, G. A. Quadir, and Z. A. Zainal, "Thermal methodology for evaluating the performance of microelectronic devices with nonuniform power dissipation," in *Proc. IEEE Electronics Packaging Technology Conf.*, 2002, pp. 312–317.

[59] J. W. Worman, "Sub-millisecond thermal impedance and steady state thermal resistance explored," in *Proc. IEEE SEMI-THERM Symp.*, 1999, pp. 173–181.

[60] Intel. Pentium Processor With MMX Technology. [Online]. Available: http://cs.mipt.ru/docs/comp/eng/hardware/processors/intel/i586/p55/main.pdf

[61] IBM . IBM 6X86MX Microprocessor. [Online]. Available: http://www-3.ibm.com/chips/techlib/techlib.nsf/techdocs/

[62] K. Banerjee and R. Mahajan. (2002) Intel Development Forum. [Online]. Available: ftp://download.intel.com/research/silicon/Thermals_press_IDF_0902.pdf

[63] H. E. Hamilton, "Thermal aspects of burn-in of high power semiconductor devices," in *IEEE Inter-Society Conf. Thermal Phenomena*, 2002, pp. 626–634.

[64] V. Szekely, "Thermal monitoring of microelectronic structures," *Microelectron. J.*, vol. 25, no. 3, pp. 157–170, 1994.

[65] B. Lian, T. Dishongh, D. Pullen, H. Yan, and J. Chen, "Flow network modeling for improving flow distribution of microelectronics burn-in oven," in *IEEE Inter-Society Conf. Thermal Phenomena*, 2000, pp. 78–81.

[66] D. Gardell, "Temperature control during test and burn-in," in *IEEE Inter-Society Conf. Thermal Phenomena*, 2002, pp. 635–643.

[67] A. Poppe, G. Farkas, M. Rencz, Z. Benedek, L. Pohl, V. Szekely, K. Torki, S. Mir, and B. Courtois, "Design issues of a multi-functional intelligent thermal test die," in *Proc. IEEE SEMI-THERM Symp.*, 2001, pp. 50–56.

[68] T. Kim and W. Kuo, "Modeling manufacturing yield and reliability," *IEEE Trans. Semiconductor Manufacturing*, vol. 12, no. 4, pp. 485–492, 1999.

[69] T. S. Barnett, A. D. Singh, M. Grady, and K. G. Purdy, "Redundancy implications for product reliability: experimental verification of an integrated yield-reliability model," in *Proc. Int. Test Conf.*, 2002, pp. 693–699.

[70] J. Van der Pol, F. Kuper, and E. Ooms, "Relation between yiled and reliability of integrated circuits and application to failure rate assessment and reduction in the one digit fit and ppm reliability era," *Microelectron. Reliabil.*, vol. 36, no. 11/12, pp. 1603–1610, 1996.

[71] T. S. Barnett and A. D. Singh, "Relating yield models to burn-in fall-out in time," in *Proc. Int. Test Conf.*, 2003, pp. 77–84.

[72] T. Kim, W. Kuo, and W.-T. K. Chien, "Burn-in effect on yield," *IEEE Trans. Electron. Packag. Manufact.*, vol. 23, pp. 293–299, Apr. 2000.

[73] A. Vassighi, O. Semenov, and M. Sachdev, "Impact of power dissipation on burn-in test environment for sub-micron technologies," in *Proc. IEEE Int. Workshop on Yield Optimization and Test*, 2001.

[74] T. Henry and T. Soo, "Burn-in elimination of a high volume microprocessor using I_{DDQ}," in *Proc. Int. Test Conf.*, 1996, pp. 242–249.

[75] R. Kawahara, O. Nakayama, and T. Kurasawa, "The effectiveness of I_{DDQ} and high voltage stress for burn-in elimination," in *Proc. IEEE I_{DDQ} Workshop*, 1996, pp. 9–14.

[76] T. Barrette, V. Bhide, K. De, M. Stover, and E. Sugasawara, "Evaluation of early failure screening methods," in *Proc. IEEE I_{DDQ} Workshop*, 1996, pp. 14–17.

[77] K. Wallquist, "On the effectiveness of I_{SSQ} testing in reducing early failure rate," in *Proc. Int. Test Conf.*, 1995, pp. 910–915.

[78] S. McEuen, "Reliability benefits of I_{DDQ}," *J. Electronic Testing: Theory and Applications (JETTA)*, vol. 3, no. 4, pp. 327–335, 1992.

[79] 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.

[80] P. Nigh, D. Vallett, P. Patel, J. Wright, F. Motika, D. Forlenza, R. Kurtulik, and W. Chong, "Failure analysis of timing and I_{DDQ}-only failures from the SEMATECH test methods experiment," in *Proc. Int. Test Conf.*, 1998, pp. 43–52.

[81] *Int. Reliability Physics Symp. (IRPS), Panel on Burn-in Elimination*, Orlando, FL, 2001.

[82] R. Daasch, K. Cota, J. McNames, and R. Madge, "Neighbor selection for variance reduction in I_{DDQ} and other parametric data," in *Proc. Int. Test Conf.*, 2001, pp. 92–100.

[83] R. Madge *et al.*, "Screening MinVDD outliers using feed-forward voltage testing," in *Proc. Int. Test Conf.*, 2002, pp. 673–682.

[84] C. Schuermyer, B. Benware, K. Cota, R. Madge, R. Daasch, and L. Ning, "Screening VDSM outliers using nominal and subthreshold supply voltage I_{DDQ}," in *Proc. Int. Test Conf.*, 2003, pp. 565–573.

Arman Vassighi recieved 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 area is 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 MOSFETs while optimizing burn-in conditions.

Oleg Semenov received the Engineer degree (with Honors) and Doctor of Science (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, ON, Canada, in 2001. His Ph.D. research was on the investigation and development of silicon-on-insulator (SOI) structures using a selective chemical etching of silicon.

He was with Joint Stock Company (Hong Kong–Russia) Korona Semiconductor, Moscow, from 1996 to 1998, where he worked as a Process Engineer. Currently, he is a Research Assistant Professor in the Department of Electrical and Computer Engineering, University of Waterloo. 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. He has contributed to more than 20 papers in various technical journals and conferences.

Manoj Sachdev (SM'97) received the B.E. degree (with Honors) in electronics and communication engineering from the 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, Agrate, Milan, Italy. In 1992, he joined Philips Research Laboratories, Eindhoven, The Netherlands, where he researched various aspects of VLSI testing and manufacturing. He is currently a Professor in the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada. His research interests include low-power and high-performance digital circuit design, mixed-signal circuit design, and 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. He holds more than 10 granted and several pending U.S. patents in the area of VLSI design and test.

Dr. Sachdev received the Best Paper Award for his paper in European Design and Test Conference, 1997, and an honorable mention award for his paper in International Test Conference, 1998.

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

He is a Staff Research Scientist with the Microprocessor Research Laboratories (MRL), Intel Corporation, Portland, OR. He is currently focusing on 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 (10 issued and 10 pending).

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

Chuck Hawkins is a Professor in the Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, 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 Corporation. He is the Editor of the *Electron Device Failure Analysis* magazine. He teaches short courses to industry with G. Hnatek on CMOS IC quality and reliability.

Dr. Hawkins is a past General and Program Chair of the International Test Conference (ITC). With co-workers at Sandia Labs, he has won several Best Paper and Honorable Mention Paper Awards at ITC and the International Symposium on Test and Failure Analysis (ISTFA). He has coauthored three books, including a nearly completed work in progress with J. Segura titled *CMOS Electronics: How it Works, How it Fails* (New York: IEEE Press, 2003).