Burn-in Temperature Projections for Deep Sub-micron Technologies

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

Electrical and Computer Engineering Dept., University of Waterloo, Waterloo, Canada N2L 3G1

*Microprocessor Research Laboratories, Intel Corporation, Hillsboro, OR 97124-6497 USA

**Electrical and Computer Engineering Dept., University of New Mexico, Albuquerque, NM 87131 USA

Abstract

Burn-in faces significant challenges in recent CMOS technologies. The self-generated heat of each IC in a burn-in environment contributes to larger currents that can lead to further increase in junction temperatures, possible thermal run away, and yield-loss of good parts. Calculations show that the junction temperature is increasing by 1.45X/generation. This paper estimates the increase in junction temperature with scaling and discusses the optimal burn-in temperature with scaling. Our research indicates that the burn-in temperature must also be reduced with technology scaling. The impact on commercial burn-in ovens is also described.

1. Introduction

Transistor scaling is the primary factor in achieving high performance microprocessors and memories. Each 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, and (3) reduced energy per transition by 65% while saving 50% of power [1-3]. To achieve this, transistor width, length, and oxide dimensions were scaled by 30%. As a result, chip area decreases by 50% for the same number of transistors, and total parasitic capacitance decreases by 30%.

The junction temperature of an IC is defined as the temperature of the silicon substrate, and it is a crucial parameter of reliability-prediction procedures and burn-in testing. 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_{DD} = 1.9 V [4]. This microprocessor had 15.2 million transistors packed in the 210 mm^2 chip area.

Scaled transistors must lower the power supply for two reasons: (1) to reduce the device internal electric fields, and (2) to reduce power consumption since power is proportional to V_{DD}^2. As V_{DD} scales, then the (V_{DD} - V_{TH}) drain current overdrive term must reduce V_{TH} to achieve higher performance. This lower V_{TH} 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 microprocessor increases as Figure 1 illustrates for Intel microprocessor power projections. Notice the increasing percentage of off-state leakage current at the 130 nm and sub-100 nm nodes. The ratio of leakage to active power becomes adverse under burn-in conditions. Typically, clock frequencies are kept in the tens of MHz range during burn-in, which results in substantial reduction in active power. On the other hand, the voltage and temperature stresses cause the off state leakage power to be the dominant power component.

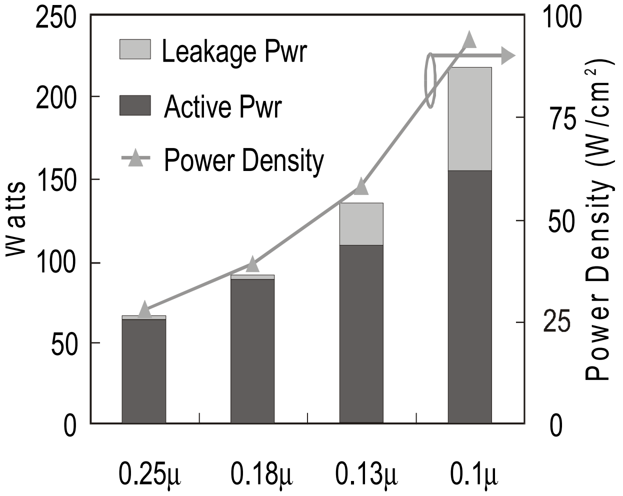

Figure 1. Power density trend adopted from [2].

Junction temperature is an estimated value taken from on-chip sensor measurements or simulation. We seek to control the junction temperature in burn-in, and the burn-in chamber temperature is the means to do so. Junction and burn-in chamber temperatures differ widely in the recent scaled nm technologies. This temperature difference presents challenging problems not seen in older technologies where burn-in chamber and package temperature were similar. This difference that we now see is caused by the IC power dissipation of all of the off-state transistors and the increased transistor density on the chips. We must actually cool the burn-in chamber to hold die junction temperature at its nominal target, i.e., 125^{o}C.

This paper develops a model and then computes the burn-in chamber temperatures to control values for different scaled technologies. The limitations for commercial burn-in chambers are analyzed with respect to burn-in in advanced technologies.

Several techniques can estimate junction temperature. One method directly measures junction temperature with thermal sensors at several on-chip locations during normal and burn-in conditions [5,6]. Another method uses chip-level 3D-electrothermal simulators that can find the steady-state CMOS VLSI chip temperature profile at the corresponding circuit performance [7,8]. However, thermal sensors are relatively large devices, and accurate prediction requires a number of them placed on the IC. Sensors require calibration. Gerosa, et. al., reported a 0.2 mm^2 thermal sensor with a sensing range of 0-128°C and a 5-bit resolution (4°C) [9]. Thermal sensors can only be used for verification, and one may have to use other techniques for prediction and estimation. 3D-electothermal simulators cannot be used for large-scale integrated circuits such as microprocessors because of large CPU time. The simulation time of a 2D Discrete Cosine Transformation (DCT) chip (107,832 transistors, 8 MHz) was reported at 12 hours [8].

We propose a method for average junction temperature (T_j) estimation that can be used for optimization of burn-in conditions. The method can predict the impact of technology scaling on burn-in conditions. We did not consider the packaging issues, such as thermal impedance of the package and other such factors. We focus on the intrinsic die behavior under the burn-in conditions, since package thermal properties tend to be user specific.

Section 2 discusses junction temperature as a parameter for reliability-prediction procedures. Section 3 explains the thermal resistance models of transistors and their impact on the junction temperature. The impact of CMOS technology scaling on average junction temperature increases during normal and burn-in conditions is analyzed in Section 4. Section 5 discusses an optimization procedure for burn-in conditions to avoid thermal runaway.

2. Junction Temperature as a Parameter of Reliability-Prediction

Accelerated tests carried out at high temperature generate reliability failures in a reasonable, short time period. We next consider several reliability-prediction models to show the importance of average junction temperature for accuracy of these procedures.

2.1 Time Dependent Dielectric Breakdown Models (TDDB) - Gate Oxide Breakdown Models

The oxide thickness decreases at each technology node to increase the drive current and to control the short channel effects. Larger drive current will charge the circuit node capacitors faster, and as a result, the circuit speed is increased. The experimental measurements of time to breakdown of ultra thin gate oxides with thickness less than 40 Å show a Voltage Driven Breakdown model [10]. The experiments show that the generation rate of stress-induced leakage current (SILC) and charge to breakdown (Q_{BD}) in ultra thin oxides is controlled by the gate voltage rather than the electric field, as described for the thicker gate oxide breakdown models. Recently, a new time to breakdown model was proposed [11]. This model (Eq. 1) includes the gate oxide thickness (T_{OX}) and the gate voltage (V_G).

$$T_{bd} = T_0 \cdot \exp\left[\gamma\left(\alpha \cdot T_{OX} + \frac{E_a}{kT_j} - V_G\right)\right] \qquad (1)$$

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/Å, γ= 12.5 1/V and $(\gamma \cdot E_a)$ = 575 meV [11].

2.2 Temperature and Voltage Acceleration Factor Models

Temperature and voltage acceleration factor models are the basis of several industrial reliability standards. The Mil-Hdbk-217F US military standard defines the temperature acceleration factor as [12]

$$\pi_T = 0.1 \exp\left(-A\left(\frac{1}{T_j} - \frac{1}{298\ K}\right)\right) \qquad (2)$$

where A is a constant and T_j is the junction temperature (K). The voltage acceleration factor is defined in the CNET reliability procedure as [13]

$$\pi_V = A_3 \exp\left[A_4 V_A\left(\frac{T_j}{298}\right)\right] \qquad (3)$$

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.

3. Semiconductor Thermal Resistance Models

The Arrhenius model predicts that the failure rate of integrated circuits is an inverse exponential function of the junction temperature. A small increase of 10-15 °C in junction temperature may result in ~2X reduction in the lifespan of the device [14]. 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 [15]

$$T_j = T + P \times R_{th} \quad (4)$$

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. We must analyze the impact of technology scaling on Eq. (4) to estimate the average junction temperature for several technologies. We investigated how the power dissipation and thermal resistance change with technology scaling in order to predict how these parameters will change.

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 [16]

$$R_{th} \approx \frac{1}{4K\,(L \times W)^{1/2}} \quad (5)$$

where K is the thermal conductivity of silicon, (L x W) is the emitter size, and R_{th} is the thermal resistance (°C/mW). 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 3-D heat flow equation [17].

$$R_{th} = \frac{1}{2\pi K}\left[\frac{1}{L}\ln\left(\frac{L+\sqrt{W^2+L^2}}{-L+\sqrt{W^2+L^2}}\right)+\frac{1}{W}\ln\left(\frac{W+\sqrt{W^2+L^2}}{-W+\sqrt{W^2+L^2}}\right)\right] \quad (6)$$

where K is the thermal conductivity of silicon (K = 1.5 x 10^{-4} W/µm°C [18]), W and L are channel geometry parameters. The thermal conductivity of silicon has a temperature dependence described in [19].

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. So, our calculations assumed that the thermal resistance of silicon was temperature independent [17,18]. We used Eq. (6) for thermal resistance calculations for MOSFETs in different CMOS technologies.

4. Scaling, Junction Temperature, Affect on Normal and Burn-in Conditions

In low-power applications, the power-supply voltage and transistor sizing are scaled more aggressively to minimize the power consumption [20, 21]. The transistor threshold voltage in low power ICs is typically higher than for high-performance ICs to suppress the sub-threshold leakage. At the same time, the speed relative to the high-performance case should not degrade more than 1.5X [20]. We will focus on high performance applications where dynamic and static power consumption are considerably high and pose a serious reliability threat.

4.1 Estimation of Junction Temperature Increase with Technology Scaling at Normal Conditions

We defined F_{max} as the maximum toggle frequency of an inverter in a given technology. The dynamic power consumption calculation under normal operating conditions was done at 70% of F_{max}. HSPICE simulations were carried out with BSIM model level 49. Transistor models for a 0.13 µm CMOS technology were taken from United Microelectronics Corporation (UMC). Transistor models for other CMOS technologies were adopted from the Taiwan Semiconductor Manufacturing Corporation (TSMC). The simulation results and transistor sizes are given in Table 1. The inverter's load was the standard load element (*n*-MOSFET) used by TSMC for inverter ring-oscillator simulations. The load element sizes were taken from the TSMC and UMC SPICE models file specified for each of analyzed CMOS technology.

Table1. Simulated CMOS inverter parameters and F_{max}.

CMOS Tech., um/V_{DD}, V	***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 x F_{max}, MHz***
0.35/3.3	*4.0/0.35*	*10.0/0.35*	*3.0/3.5*	*1450*	*1015*
0.25/2.5	*2.86/0.25*	*7.14/0.25*	*2.15/2.5*	*1950*	*1365*
0.18/1.8	*2.06/0.18*	*5.14/0.18*	*1.55/1.8*	*2300*	*1610*
0.13/1.2	*1.49/0.13*	*3.71/0.13*	*1.12/1.3*	*4000*	*2800*

The International Technology Roadmap for Semiconductors (ITRS) 2002 [22] indicates that the scale-down of device size is still in progress. The planar type transistors with 15-30 nm gate length have already been demonstrated [23]. However, a 90-100 nm CMOS technology is considered today as a next generation for microprocessor and SRAM chips [24-26]. Therefore, we included the 90 nm CMOS technology node in our study of burn-in testing. The effective channel length of transistors for this technology is assumed at 55-65 nm.

We simulated the total power consumption of an inverter toggling at 0.7 F_{max} in four different technologies with results given in Table 1. The thermal resistance of an average transistor was computed from Eq. (6). The average size of a transistor was estimated by averaging the *n*MOS and *p*MOS transistor widths. Since, the transistor dimensions were reduced, the thermal resistance increased with scaling. Figure 2 illustrates inverter power dissipation at an operating frequency of

0.7 F_{max} and thermal resistance of an average transistor as functions of technology.

Owing to lack of technology access for 90 nm CMOS technology, an alternative method was utilized for inverter power and thermal resistance estimations in Figure 2. For the 1.0 V, 90 nm CMOS technology, the ITRS predicts the transistor density in a microprocessor chip to be about 0.27 millions/mm^2. It is assumed that the transistor density is doubled with technology scaling for each new process generation. The industrial estimation of power density of a microprocessor chip, implemented in 90 nm technology, is approximately 0.5 W/mm^2 [1,25,26]. Power density is defined as the power dissipated by the chip per unit area under nominal frequency and normal operating conditions. Using these assumptions we can estimate the inverter power dissipation at normal operating conditions (V_{DD} = 1 V, T = 25 °C) and speed (Figure 2).

We also extended the scaling scenario of transistor sizes in a CMOS inverter to 90 nm CMOS technology to calculate the thermal resistance. We used transistor sizes of P-MOSFET (W/L)=3.0/0.1 and N-MOSFET (W/L)=1.0/0.1. The calculated transistor thermal resistance for 90 nm technology using Eq. (6) is shown in Figure 2.

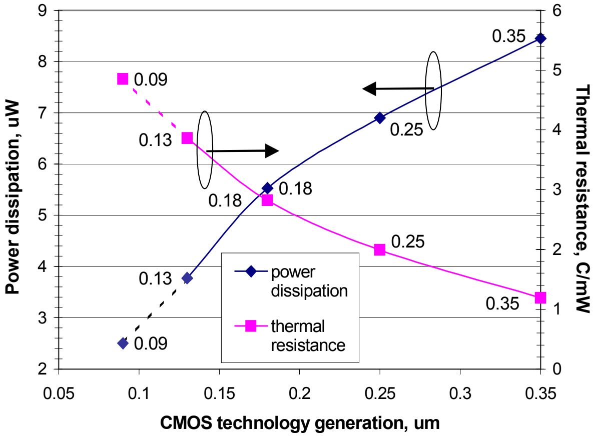

Figure 2. Inverter power dissipation and transistor thermal resistance for different CMOS technologies.

We used the 0.35 µm CMOS technology as the reference technology. Eq. (4) defines ΔT as the temperature difference between junction and the ambient. If ΔT is unity for a 0.35 µm technology, then we may calculate the normalized change in ΔT with respect to the reference technology. Using Eq. (4) and data presented in Figure 2, we estimated the normalized average temperature increase for different technologies. For example, Eq. (7) is used for calculation of $\Delta T_{0.25}/\Delta T_{.0.35}$ ratio:

$$\frac{\Delta T_{0.25}}{\Delta T_{0.35}} = \frac{(T_j - T)_{0.25}}{(T_j - T)_{0.35}} = \frac{(P \times R_{th})_{0.25}}{(P \times R_{th})_{0.35}} \quad (7)$$

Figure 3 shows the normalized MOSFET junction temperature change with respect to the 0.35 µm technology using Eq. (7). As the technology went from 0.35 µm to 0.18 µm, the normalized temperature increased primarily from the increase in thermal resistance with scaling. However, scaling from 0.18 µm to 0.09 µm resulted in lower normalized MOSFET junction temperature with respect to 0.18 µm technology. The reduction in normalized transistor temperature is due to the drastic reduction in power dissipation. The reduced parasitic capacitance is the primary reason in reduced power dissipation.

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) [22,27]. Figure 4 shows 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.

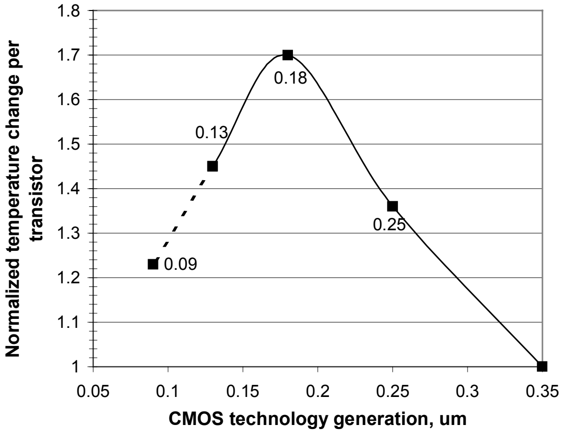

Figure 3. MOSFET junction temperature vs. technology.

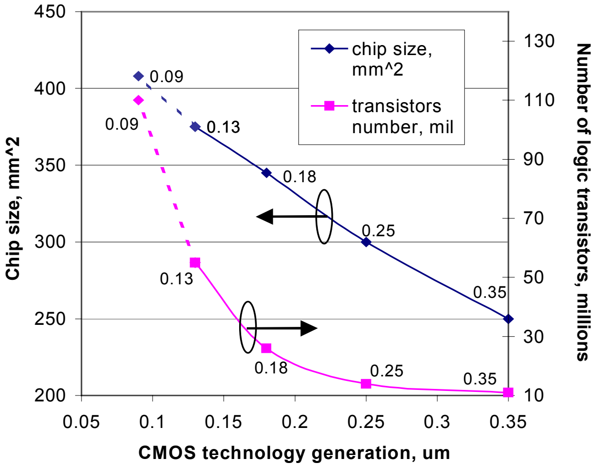

Figure 4. The trends of CMOS logic chips (data for graphs were adopted from [22,27]).

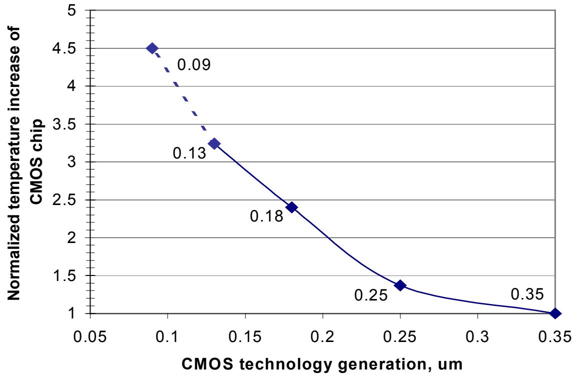

Figure 5. Normalized chip junction temperature increase with technology.

The normalized temperature increase of a CMOS chip with technology scaling was calculated by multiplying the temperature increase per transistor in Figure 3 times the transistor density calculated from Figure 4. The results are shown in Figure 5. We conclude from Figure 5 that the normalized temperature increase of the chip is significantly elevated with CMOS technology scaling from 350 nm 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 0.35 µm CMOS chip. This calculation assumed that the ambient temperature was the same for all analyzed technologies. The increase in chip junction temperature results in an exponential increase in cooling cost [6].

4.2 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 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 life-time 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. 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. For the DLBI, one must also consider the thermal impedance network of the package [28]. Once this network is known, then Eq. (6) can be suitably modified to reflect the total thermal resistance (R_{th}) of the die and many types of package. In this article, we focus on the intrinsic behavior (junction temperature estimation) of the silicon die under burn-in conditions for the sake of simplicity. In other words, the thermal impedance network of the package is not considered.

We estimated the average inverter power for different operating conditions and technologies (Table 1) by simulating the inverters at different temperatures and V_{DD}. For burn-in, we varied the stress temperature from 25 °C to 125 °C. Similarly, the stress voltage was varied from nominal V_{DD} for the given technology to V_{DD} + 30%, and in this simulation (BSIM model level 49) the inverter input was grounded. The simulated I_{av} and the calculated values of P and ΔT are given in Table 2, where I_{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 Eq. (8).

$$\Delta T = P_{\text{transistor}} \times R_{th\text{-transistor}} \times \frac{D_{\text{density}}}{2} \quad \left[\frac{^{\circ}C}{mm^{2}}\right] \qquad (8)$$

where $P_{transistor}$ is the power dissipation of the off-mode transistor in the inverter, $R_{th\text{-}transistor}$ is the thermal resistance of the on-transistor in the inverter, and $D_{density}$ is the transistor density in the CMOS chip. For a given technology, the thermal resistance was extracted from Figure 2 and the transistor density was calculated from Figure 4. We assumed a fully static CMOS design. Therefore, half of the total transistors are in the off-mode during burn-in, and this was taken into account by dividing $D_{density}$ by 2 in Eq. (8).

Since we did not have access to industrial HSPICE device models for the 90 nm CMOS technology, we could not use HSPICE simulations in Cadence for this technology generation. To predict the possible increase of average junction temperature in CMOS chips under burn-in conditions, we simulated an *n*MOSFET at stressed operating conditions using a 2-D device simulator "Microtec" [29]. The MOSFET parameters used for device simulations are given in Table 3. The simulation results correspond to DC characteristics of 90 nm transistors [24,25,30], such as V_{TH} = 0.2-0.28 V, I_{ON} = 600-750 uA/um and I_{OFF} = 20-100 nA/um. These devices were developed for ultra high performance applications (UHP). Low power (LP) medium speed [24,30] devices assume V_{TH} = 0.3-0.35 V, I_{ON} = 480-520 uA/um and I_{OFF} = 0.18-0.5 nA/um. High performance (HP) applications assume a leakage current of approximately 10 nA/um [3].

In this section we consider UHP and LP devices as worst and best cases with respect to power consumption during burn-in. The transistor parameters obtained from simulations under normal operating conditions are presented in Table 4. The dominant components of leakage current of a sub-100 nm MOSFET are sub-threshold, band-to-band tunneling, and gate oxide tunneling currents [26].

The simulation results of averaged size MOSFET (W/L=2.0 µm/0.1 µm) under stressed operating conditions are given in Table 5. In this table, P is the power dissipation of an off-mode inverter transistor that was obtained from device simulations. ΔT is the (T_j - T) per 1 mm^2 of CMOS chip that was calculated by Eq. (8). The transistor density in CMOS chip was assumed to be 0.27 millions/mm^2 (Figure 4). When we calculated ΔT in Table 2 and Table 5, we considered that each off-mode transistor in a 1 mm^2 chip area was an independent heat source. 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, and this depends on layout. In the first order approximation, we neglected the thermal coupling effect of transistors in our analysis. Table 2 and Table 5 show 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. At 125°C, the increase in current and power dissipation with technology scaling is relatively less. However, the increase in ΔT is more dramatic owing to increased transistor density, leakage current, and the thermal resistance.

The normalized temperature increase of a CMOS chip with scaling at burn-in conditions is shown in Figure 6. The graph with diamond legend depicts the normalized T_j increase if T = 125°C. For 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_{TH} UHP devices (unrealistic) then the normalized T_j is increased by approximately 5000x compared to 0.35 µm CMOS. On the other hand, if all the transistors are implemented with LP devices, then the T_j is increased by approximately 230x.

Table 3. n-MOSFET parameters used for simulations.

	UHP	LP
Substrate doping, cm^{-3} (p - type)	5×10^{15}	5×10^{15}
Source/Drain doping, cm^{-3} (n - type)	3×10^{20}	3×10^{20}
V_{TH} adjusted doping, cm^{-3} (p - type)	1.8×10^{18}	3×10^{18}
Punch - Trough doping, cm^{-3} (p - type)	5×10^{19}	8×10^{19}
Effective gate oxide thickness, Å	18	18
L_{eff}/W, nm/µm	63/2	63/2
Nominal V_{DS}=V_{DD}, V	1.0	1.0

Table 4. DC parameters of n-MOSFET emulated in 90 nm CMOS technology (V_{DD} = 1V, T = 25 °C).

	L_{eff}, nm	V_{TH}, V	I_{ON}, uA/um	I_{OFF}, nA/um
UHP	63	0.25	600	30
LP	63	0.35	440	0.6

Table 2. DC simulation (I_{av}) and calculation results (P, ΔT) of CMOS inverters for different technologies.

CMOS technology		25 °C			85 °C			125 °C		
		I_{av}, pA	P, pW	ΔT, °C/mm^2	I_{av}, nA	P, nW	ΔT, °C/mm^2	I_{av}, nA	P, nW	ΔT, °C/mm^2
0.35 -um	V_{DD} = 3.3 V	7.7	25	0.00071	0.07	0.23	0.0066	2.05	6.77	0.2
	V_{DD} = 3.8 V	9.2	35	0.00099	0.084	0.32	0.0091	2.15	8.17	0.23
	V_{DD} = 4.3 V	11.1	47.7	0.0014	0.11	0.47	0.014	2.27	9.76	0.28
0.25 -um	V_{DD} = 2.5 V	19.3	48.3	0.0023	0.418	1.04	0.05	3.96	9.9	0.29
	V_{DD} = 2.9 V	22	63.8	0.0031	0.47	1.36	0.065	4.41	12.80	0.35
	V_{DD} = 3.25 V	25	81.3	0.0039	0.531	1.75	0.08	4.81	15.87	0.45
0.18 -um	V_{DD} = 1.8 V	90.5	163	0.02	1.33	2.39	0.24	8.96	16.13	0.97
	V_{DD} = 2.1 V	101	210	0.022	1.48	3.08	0.31	9.75	20.48	1.23
	V_{DD} = 2.35 V	112	264	0.027	1.62	3.81	0.39	10.9	25.6	1.51
0.13 -um	V_{DD} = 1.2 V	766	920	0.2	8.45	10	2.32	28	34	7.79
	V_{DD} = 1.4 V	1200	1680	0.38	12.3	17	3.94	34	47	10.97
	V_{DD} = 1.56 V	1860	2900	0.67	17.7	27.6	6.4	55	85	19.81

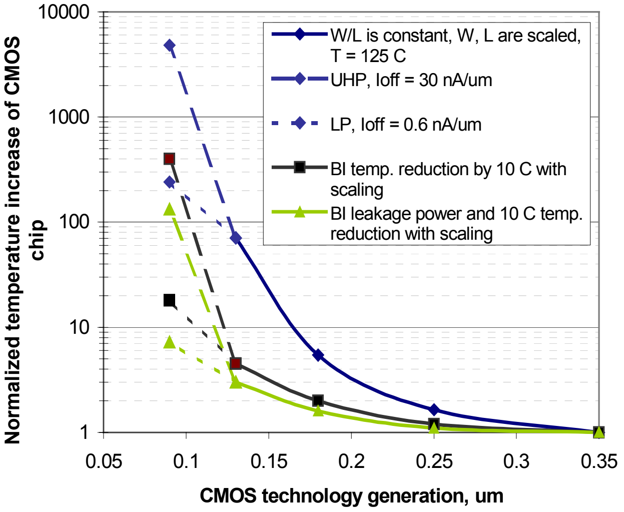

Figure 6. Normalized chip junction temperature at V_{DD} + 30% burn-in condition.

It should be noted that most of the transistors on chip will be implemented with LP devices.

However, if the T is reduced by 10°C for each technology generation the normalized T_j is also reduced as shown by the graph with square legend. Similarly, leakage reduction techniques can also be employed to further reduce the increased normalized temperature with scaling [31,32]. If such techniques are employed as well as the T is reduced by 10°C for each technology generation the normalized T_j increase for 90 nm CMOS with respect to 0.35 µm CMOS becomes relatively small (7-8x).

In spite of 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.

5. Burn-in Limitations and Optimization to Avoid Thermal Runaway

Several reliability failure mechanisms are accelerated by temperature, so burn-in testing is done at elevated temperature. These mechanisms include metal stress voiding and electromigration, metal slivers bridging shorts, contamination, and gate-oxide wearout and breakdown [33]. 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 [34]. As a result, the yield loss may increase if the burn-in conditions are overstressed. Hence, we should optimize the junction temperature of die for normal and burn-in conditions.

5.1 Physical and Practical Limits of Junction Temperature

Table 5. Predicted power dissipation and junction temperature increase of CMOS inverter (90 nm CMOS technology).

		-100 °C		0 °C		25 °C		85 °C		125 °C	
	V_{DD}, V	P, nW	ΔT, °C/mm²	P, nW	ΔT, °C/mm²	P, nW	ΔT, °C/mm²	P, nW	ΔT, °C/mm²	P, nW	ΔT, °C/mm²
UHP	**1.0**	0.057	0.021	18.4	8.03	60	26.2	328	143.2	1344	586.8
	1.15	0.084	0.036	29	12.66	82.8	36.15	506	221	2047	893.8
	1.3	0.120	0.052	44.2	19.3	130	56.8	770	336.2	3084	1346.6
		-100 °C		**0 °C**		**25 °C**		**85 °C**		**125 °C**	
	V_{DD}, V	P, pW	ΔT, °C/mm²	P, nW	ΔT, °C/mm²	P, nW	ΔT, °C/mm²	P, nW	ΔT, °C/mm²	P, nW	ΔT, °C/mm²
LP	**1.0**	0.124	$5.4x10^{-5}$	0.32	0.14	1.2	0.53	9.8	4.28	74.4	32.49
	1.15	0.208	$9.1x10^{-5}$	0.51	0.23	1.84	0.81	14.63	6.39	107.6	46.99
	1.3	0.34	$1.5x10^{-4}$	0.75	0.33	2.7	1.18	21.23	9.27	152.9	66.75

The maximum operating temperatures for semiconductor devices can be estimated from the semiconductor intrinsic carrier density that 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 reported operating junction temperature is about 200 °C in standard silicon technology [35]. At this temperature, the circuit performance is reduced substantially. The temperature will affect thermal conductivity, built-in potential, threshold voltage, and *pn* junction reverse current. Several practical considerations limit the junction temperature to a much lower value. A value of 150 °C for junction temperature is often used for ICs as the limit [22].

The peak junction temperature of a PowerPC microprocessor implemented in a 0.35 µm CMOS technology with a 0.3 µm effective transistor channel lengths is about 90°C - 100°C at an operating speed of 200 - 250 MHz [36,9]. If we use this as the reference temperature and assume that Figure 5 estimates the junction temperature increase with reasonable accuracy, then we should expect a 2.4X increase in junction temperature for the same microprocessor implemented in a 0.18 µm CMOS technology. Hence, the die junction temperature should be approximately 156°C - 180°C.

These values are obtained assuming cooling, packaging and circuit techniques remain the same when moving from 0.35 µm technology to 0.18 µm technology. However, improved cooling and packaging considerations will reduce the temperature to much lower value. Similarly, circuit techniques such as transistor stacking, dual-threshold transistors, reverse body bias, etc. can reduce substantially leakage current and the junction temperature.

5.2 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 Watts [37]. If we use the power dissipation of a single transistor in an inverter at static stressed conditions from Table 2 and Table 5, and use the number of transistors of the logic chip for different CMOS technologies from Figure 4, then we can estimate the maximum number of die for different technologies that can be simultaneously powered in a burn-in oven using Eq. (9).

$$N_{dies} = \frac{P_{oven}}{P_{transistor} \times \frac{N_{transistors}}{2}} \qquad (9)$$

where P_{oven} is the maximum power dissipation of the burn-in oven at stressed conditions, $P_{transistor}$ is the power dissipation of a single transistor at static stressed conditions for the given technology, and $N_{transistors}$ is the total number of transistors in the logic chip for the given technology. Eq. (9) assumes that 50% of the total number of transistors are off at any point during burn-in assuming fully static CMOS design.

Burn-in ovens, such as the PBC1-80 of Despatch Industries [37] and Max-4 of Aehr Test Systems [38] have maximum power dissipation of about 2500 and 15,000 watts respectively at 125°C. The room ambient temperature is assumed to be 25°C. Now from Eq. (9), we can calculate the maximum number of die that can be powered during burn-in. Figure 7 plots this calculation over several technology generations. This figure shows that the maximum number of die that can simultaneously be powered in burn-in is reduced. If the burn-in temperature is kept constant while stress voltage is correspondingly reduced (T=125 °C, V_{DD}+30%), the corresponding graph depicts an exponential reduction in number of dies that can be burnt together. The elevated leakage with scaling is the main reason in this reduction. If the stress temperature is reduced by 10 °C for each technology generation while stress voltage is kept at V_{DD}+30%, the reduction in dies is modest with scaling. A larger number of dies can be burnt-in together, if the chip has leakage reduction techniques, as well as, the temperature is reduced by 10 °C for each technology generation.

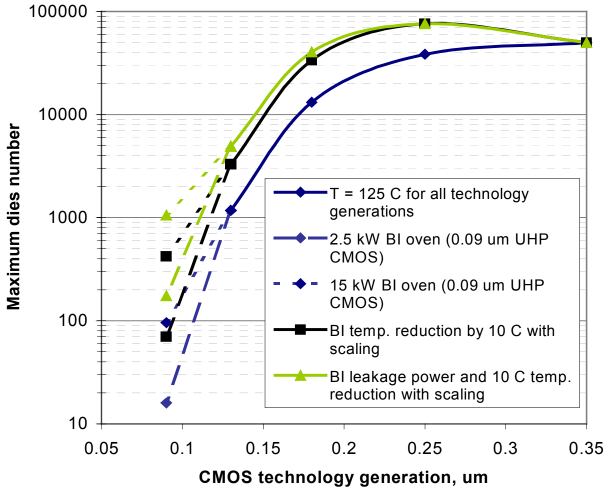

Figure 7. Maximum dies number for one burn-in load versus CMOS technology scaling.

5.3 Optimization of Burn-In Stress Conditions with Technology Scaling for Constant Reliability

The optimal burn-in conditions for maintaining the projected failure rate require that the defect distribution models and their growth models be studied. The post burn-in reliability (R) and yield loss (Y_{loss}) were studied by researchers [39,40]. T. Kim, et al., [40] proposed the following models for post burn-in reliability and yield loss shown in Eq. (10) and Eq. (11).

$$Y_{loss} = Y\left(1 - Y^{\frac{\upsilon}{1-\upsilon}}\right) \qquad (10)$$

$$R = Y^{\frac{1}{(1-u)^{2}-1}} \qquad (11)$$

where Y is the yield before burn-in, and υ, u are constants that depends on stress temperature and voltage. Using the 1/E gate oxide breakdown model and the post burn-in yield loss model, Vassighi, et al., demonstrated that the post burn-in yield loss is increased exponentially with the elevation of stress temperature for a given stress voltage [39]. This result was obtained for a 0.18 µm CMOS technology ($T_{OX} \approx 41$ Å).

Hence, over-stressed die during burn-in may significantly reduce the post burn-in reliability and increase the yield loss, especially when junction temperature at burn-in and normal operating conditions are increased with technology scaling. Thus, to a first order, we want a constant reliability during burn-in 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 reliability for an advanced CMOS technology should not be worse than the post burn-in reliability for the 0.35 µm CMOS technology. This means that the junction temperature increase over ambient temperature during burn-in for subsequent technologies should not be higher than the burn-in junction temperature increase for 0.35 µm CMOS technology. Table 2 shows that for 0.35 µm CMOS technology, the junction temperature increase (ΔT) over ambient stressed temperature per mm^2 of chip is 0.28°C at V_{DD} = 4.3 V, T = 125°C. The horizontal line on Figure 8 illustrates this limit. Now for 0.25 µm technology, if we plot $\Delta T/mm^2$ versus stress temperature for three different stress voltages, it results in three different curves. Subsequently, we find the optimal burn-in temperature where the horizontal line ($\Delta T = 0.28°C/mm^2$) intersects with graphs.

Similarly, we can find the optimal burn-in temperature for other technologies using data from Table 2 and Table 5. The results are shown in Figure 9 where the optimal burn-in temperature is presented for different technologies. Squares represent the data points for each technology. In this figure, voltage is kept at V_{DD} + 30% for each technology. These data points were plotted ensuring that the average junction temperature increase over ambient (ΔT) of die for these technologies is the same as the average ΔT increase for a 0.35 µm CMOS technology. Hence, we expect that the post burn-in reliability for scaled CMOS technologies has the same value as the post burn-in reliability for 0.35 µm CMOS technology.

Figure 9 shows that the optimal burn-in temperature is reduced with scaling. This observation is in line with the recently presented data showing for 0.18 µm microprocessor, the burn-in temperature is 85°C - 90°C [41]. As mentioned before, if leakage reduction techniques are employed (diamond data points), the optimal burn-in temperature is increased for 0.18 µm or lower geometries. For example, according to our research, the optimal temperature for 130 nm technology ($V_{DD} \approx 1.4$ V) is approximately 10 °C (without leakage reduction techniques) and 35°C (with leakage reduction techniques).

Furthermore, 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 reliability worse than that of the 0.35 µm CMOS technology. For example, the estimated burn-in temperature for a 90 nm CMOS technology may be approximately 0 °C to 15 °C. Note, that most of chips will use mixture technology: UHP logic is for critical delay paths and LP logic is for the low-activity SRAM cells [26].

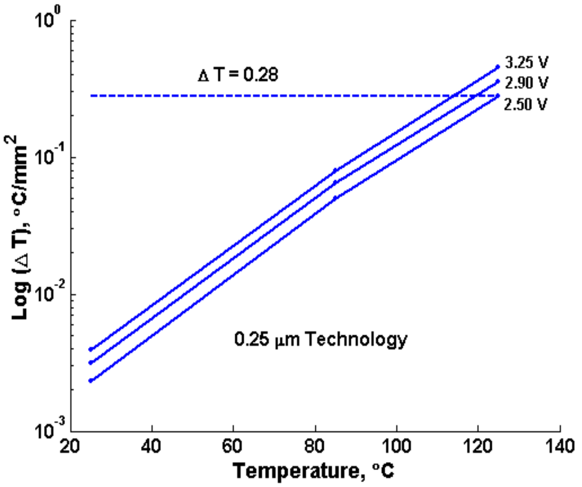

Figure 8. ΔT as a function of ambient temperature and V_{DD} for 0.25 µm technology.

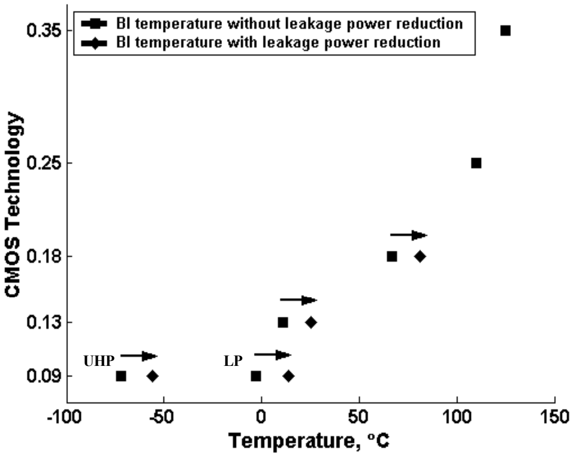

Figure 9. Optimal burn-in temperature for constant reliability.

6. Conclusion

We investigated the impact of technology scaling on the burn-in environment. The following conclusions are obtained:

Firstly, there is a steady increase in junction temperature with scaling. Under normal operating conditions, the normalized increase in junction temperature is estimated as 1.45X/generation. Similarly, the normalized junction temperature increase under burn-in conditions becomes exponential with technology scaling if no leakage reduction techniques are used. On the other hand, if leakage reduction techniques are used, it results in approximately linear increase in junction temperature. As a consequence, the burn-in temperature must be reduced with scaling.

Secondly, the number of dies that can be simultaneously burnt-in is reduced with the technology scaling, because of the maximum power dissipation limit of burn-in ovens.

Finally, the optimal stressed temperature in a burn-in environment is significantly reduced with technology scaling.

Acknowledgement

Authors thank T.M. Mak from Intel Corporation for fruitful discussions on burn-in, and Gene Hnatek (Qual Comm), Steve Steps (Aehr Test Systems), and Bill Mann for critical comments.

References

[1] S. Borkar, "Design challenges of technology scaling," IEEE Micro, pp. 23-29, July-August, 1999.

[2] S. Rusu, "Trends and challenges in VLSI technology scaling toward 100 nm," Presentation in ESSCIRC 2001, web-page: http://www.esscirc.org/esscirc2001/C01_Presentati.ns/404.pdf

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

[4] J. Ahn, H.-S. Kim, T.-J. Kim, H.-H. Shin, Y.-Ho Kim, D.-U. Lim; J. Kim, U. Chung; S.-C. Lee, K.-P. Suh, "1GHz microprocessor integration with high performance transistor and low RC delay," Proc. of IEDM, 1999, pp. 28.5.1 - 28.5.4.

[5] T.J. Goh, A.N. Amir, C.-P. Chiu, J. Torresola, "Novel thermal validation metrology based on non-uniform power distribution for Pentium III Xeon cartridge processor design with integrated level two cache," Proc. of Electronic Components and Technology Conference, 2001, pp. 1181 -1186.

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

[7] Y.K. Cheng, C.-C. Teng, A. Dharchoudhury, E. Rosenbaum, S.-M. Kang, "A chip-level electrothermal simulator for temperature profile estimation of CMOS VLSI chips," Proc. of Int. Symposium on Circuit and Systems, 1996, pp. 580-583.

[8] C.-C. Teng, Y.-K. Cheng, E. Rosenbaum, S.-M. Kang, "iTEM: A temperature-dependent electromigration reliability diagnosis tool," IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 16, No. 8, pp. 882-893, 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, G. Gerosa, "A 250-MHz 5-W PowerPC microprocessor with on-chip L2 cash controller," IEEE J. of Solid-State Circuits, vol. 32, No. 11, pp. 1635-1649, 1997.

[10] P.E. Nicollian, W.R. Hunter and J.C. Hu, "Experimental evidence for voltage driven breakdown models in ultra thin gate oxides," Proc. of IEEE Int. Reliability Physics Symposium, 2000, p.7-15.

[11] F. Monsieur, E. Vincent, D. Roy, S. Bruyere, G. Pananakakis, G. Ghibaudo, "Time to breakdown and voltage to breakdown modeling for ultra-thin oxides (T_{OX}<32 Å)," Proc. of IEEE Int. Reliability Workshop (IRW) , 2001, p.20-25.

[12] P. Lall, "Tutorial: Temperature as an input to microelectronics-reliability models," IEEE Trans. on Reliability, vol.45, No. 1, pp, 3-9, 1996.

[13] J.B. Bowles, "A survey of reliability-prediction procedures for microelectronics devices," IEEE Trans. on Reliability, vol.41, No. 1, pp, 2-12, 1992.

[14] R. Viswanath, V. Wakharkar, A. Watwe, V. Lebonheur, "Thermal performance challenges from silicon to systems," Intel Tech. Journal, Q3, pp. 1-16, 2000. http://developer.intel.com/technology/itj/archive.htm

[15] P. Tadayon, "Thermal challenges during microprocessor testing," Intel Techn. Journal, Q3, pp. 1-8, 2000. http://developer.intel.com/technology/itj/archive.htm

[16] R.C. Joy and E.S. Schlig, "Thermal properties of very fast transistors," IEEE Trans. on Electron Devices, vol. ED-17, No. 8, pp. 586-594, 1970.

[17] N. Rinaldi, "Thermal analysis of solid-state devices and circuits: an analytical approach," Solid-State Electronics, vol.44, No.10, pp. 1789-1798, 2000.

[18] N. Rinaldi, "On the modeling of the transient thermal behavior of semiconductor devices," IEEE Trans. on Electron Devices, vol. 48, No. 12, pp. 2796-2802, 2001.

[19] D.L. Blackburn and A.R. Hefner, "Thermal components models for electro-thermal network simulation," Proc. of 9th IEEE SEMI-THERM Symposium, 1993, pp. 88-98.

[20] B. Davari, R.H. Dennard, G.G. Shahidi, "CMOS scaling for high performance and low power - The next ten years," Proc. of the IEEE, vol. 83, No. 4, pp. 595-606, 1995.

[21] M. Borah, R.M. Owens, M.J. Irwin, "Transistor sizing for low power CMOS circuits," IEEE Trans. on Computer-Aided Design of Int. Circuits and Systems, vol. 15, No 6, pp. 665-671, 1996.

[22] International Technology Roadmap for Semiconductors (ITRS), http://public.itrs.net/

[23] B. Doyle, R. Arghavani, D. Barlage, S. Datta, M. Doczy, J. Kavalieros, A. Murthy, R. Chau, "Transistor element for 30 nm physical gate lengths and beyond," Intel Tech. Journal, vol. 6, No. 2, pp. 42-54, 2002.

[24] S.-F. Huang, C.-Y. Lin, Y.-S. Huang, T. Schafbauer, M. Eller, Y.-C. Cheng, S.-M. Cheng, S. Sportouch, W. Jin, N. Rovedo, A. Grassmann, Y. Huang, J. Brighten, C.H. Liu, B. von Ehrenwall, N. Chen, J. Chen; O.-S. Park, M. Common, "High-performance 50 nm CMOS devices for microprocessors and embedded processor core applications," IEDM, 2001, pp. 237-240.

[25] A. Ono, K. Fukasaku, T. Hirai, S. Koyama, M. Makabe, T. Matsuda, M. Takimoto, Y. Kunimune, N. Ikezawa, Y. Yamada, F. Koba, K. Imai, N. Nakamura, "A 100 nm node CMOS technology for practical SOC application requirement," IEDM, 2001, pp. 511-514.

[26] D.J. Frank, "Power-constrained CMOS scaling limits," IBM J. Research & Development, vol. 46, No 2/3, pp. 235-244, 2002.

[27] D.P. Valett and J.M. Soden, "Finding fault with deep-submicron ICs," IEEE Spectrum, pp. 39-50, October 1997.

[28] G. Kromann, "Thermal management of a C4/CBGA interconnect technology for a high-performance RISC microprocessor: The Motorola PowerPC 620^{TM} microprocessor," Proc. of IEEE Electronic and Tech. Conf., pp. 652-659, 1996.

[29] Siborg Corp. web-site: http://www.siborg.com/

[30] K. Fukasaku, A. Ono, T. Hirai, Y. Yasuda, N. Okada, S. Koyama, T. Tamura, Y. Yamada, T. Nakata, M. Yamana, N. Ikezawa, T. Matsuda, K. Arita, H. Nambu, A. Nishizawa, K. Nakabeppu, N. Nakamura, "UX6-100 nm generation CMOS integration technology with Cu/Low-k interconnect," Proc. of Symposium on VLSI Technology, 2002, pp. 64 -65.

[31] S. Borkar, "Leakage reduction in digital CMOS circuits," Proc. of IEEE Solide-State Circuits Conf., pp. 577-580, 2002.

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

[33] Righter A.W., Hawkins C.F., Soden J.M., Maxwell P., "CMOS IC reliability indicators and burn-in economics," Proc. of Int. Test Conf., 1998, pp. 194-203.

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

[35] W. Wondrak, "Physical limits and lifetime limitations of semiconductor devices at high temperature," Microelectronics Reliability, vol. 39, No. 6-7, pp. 1113-1120, 1999.

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

[37] Despatch Industries, http://www.despatch.com/pdfs/PBC.pdf

[38] Aehr Test Systems, http://www.aehr.com/

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

[40] T. Kim, W. Kuo, Wei-Ting Kary Chien, "Burn-in effect on yield," IEEE Trans. on Electronics Packaging Manufacturing, vol. 23, No. 4, pp. 293-299, 2000.

[41] T.M. Mak, "Is CMOS more reliable with scaling?" IEEE Int. On-Line Testing Workshop, July, 2002