Analysis of Temporal and Spatial Temperature Gradients for IC Reliability

Zhijian Lu[†], Wei Huang[†], Shougata Ghosh[†], John Lach[†], Mircea Stan[†], Kevin Skadron[‡]

Departments of [†]Electrical and Computer Engineering and [‡]Computer Science, University of Virginia

Charlottesville, VA 22904

Abstract

One of the most common causes of IC failure is interconnect electromigration (EM), which exhibits a rate that is exponentially dependent on temperature. As a result, EM rate is one of the major determinants of the maximum tolerable operating temperature for an IC and of resulting cooling costs. Previous EM models have assumed a uniform, typically worst-case, temperature. This paper presents a model that accounts for temporal and spatial variations in temperature, and shows that accounting for these variations can dramatically improve interconnect lifetime prediction accuracy. We also show that the same modeling approach applies to temperature-related gate-oxide breakdown, another common cause of IC failure. We then propose that by modeling expected lifetime as a resource that is consumed over time at a rate dependent on temperature, substantial design margin can be reclaimed. For example, for a fixed target lifetime, intermittent higher operating temperatures and performance can be tolerated if compensated by lower temperatures at other times during the product's lifetime. This approach offers higher overall performance and/or lower cooling costs than a standard design methodology that uses a worst-case temperature criterion for reliability analysis. *This report supersedes TR CS-2003-21 and TR CS-2004-07.*

Keywords—reliability, electromigration, gate oxide breakdown, thermal management, gradient

I. Introduction

As CMOS technology continues to scale, power density of VLSI circuits has unfortunately been scaling too, and this rapid increase is widely expected to continue. Yet tolerable operating temperatures usually remain fixed from generation to generation, independent of power density. This is because many aging mechanisms in VLSI circuits proceed at a rate that is temperature dependent, and product-lifetime requirements dictate the maximum aging rate. For most applications, maximum temperatures are therefore limited to around 100–120° C. Unfortunately, this means that rising power densities impose rising cooling costs that may eventually become so prohibitive that they limit the development of new products.

Historically, the temperature dependence of many aging processes can be empirically modeled by the Arrhenius Equation

$$\text{MTF} = \text{MTF}_0 \, exp\left(\frac{E_a}{kT}\right) \tag{1}$$

where MTF_0 is the mean time to failure at a specified reference temperature, E_a is the activation energy of the failure, and k is the Boltzmann constant. A detailed model based on the physics of interconnect electromigration (EM) and gate-oxide breakdown, two common temperature-dependent IC aging processes, can be found in [2] and [1].

However, these models do not account for dynamic temporal or spatial temperature variations, which have been experimentally shown to have a significant impact on circuit lifetime [5]. Using existing models, circuit designers must assume a constant temperature (usually the worst-case temperature) for the entire circuit, resulting in inaccurate lifetime estimations and excessive design margins.

This paper presents a temperature-dependent reliability model for interconnect EM. The most important contribution in this model is the ability to take into account temporal and spatial temperature gradients and, therefore, more accurately predict circuit lifetime. We also show that the same modeling approach can be used to account for temporal temperature gradients when modeling gate-oxide breakdown. Compared to a standard design methodology that uses a worst-case temperature threshold to achieve an expected lifetime, reliability models that account for temporal and spatial gradients are able to extract higher performance given a particular cost constraint, or to reduce cooling costs while maintaining performance. A particularly attractive approach for taking advantage of information about temperature gradients is to treat circuit lifetime as a resource that is consumed over time by temperature, which fits well with recent techniques for *dynamic*, runtime thermal-management techniques that adapt to specific workload behavior to maximize circuit performance while still meeting reliability requirements. A chip can, in real time, use periods of low temperature to offset brief periods of higher temperature.

II. Interconnect Reliability Analysis With Temperature Gradients

Interconnect EM is the process of self-diffusion due to momentum exchange between electrons and metal atoms. Clement [2] presents a 1-D analytical model of EM-induced interconnect stress build-up that matches empirical measurements. An interconnect failure occurs when the stress reaches a threshold value σ_{th}. The stress build-up caused by atom dislocation can be described by the following equation [2]

$$\frac{\partial \sigma}{\partial t} - \frac{\partial}{\partial x}\left[D_a\left(\frac{B\Omega}{kTL^2\varepsilon}\right)\left(\frac{\partial \sigma}{\partial x} - \frac{q^*LE}{\Omega}\right)\right] = 0 \tag{2}$$
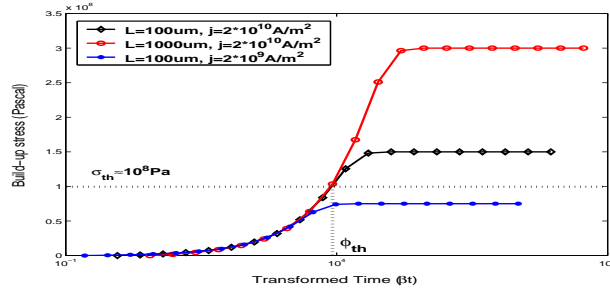
Fig. 1. Numerical solutions of Eq.(4) follow the same track before failure for different values of β.

where $\sigma(x,t)$ is the stress function, x is the 1-D position and has been normalized to $[-1,0]$ for simplicity; D_a is the diffusivity of metal atoms, which has the Arrhenius form of temperature dependency: $D_a = D_{a0}\,exp(-Q/kT)$; B, Ω and ε are constants depending on the properties of the metal, the surrounding materials and the interconnect aspect ratio; L is the characteristic length of the interconnect; q^* is the effective charge; and E is the applied electric field, which is equal to ρj, the product of resistivity and current density. The term q^*LE/Ω corresponds to a back-flow flux due to the electric field and determines the steady-state stress solution of Eq.(2).

Clement's reliability analysis was performed assuming a temporally and spatially uniform temperature [2]. In the following two subsections, we extend his model to investigate the effects of temporal and spatial gradients to interconnect lifetime predictions.

A. Temporal temperature gradients

From Eq.(2), if we define

$$\beta(T) = D_a \frac{B\Omega}{kTL^2\varepsilon} = D_{a0}\frac{B\Omega}{kTL^2\varepsilon}\,exp(-Q/kT) \tag{3}$$

and

$$\alpha(j) = \frac{q^*\rho jL}{\Omega} \tag{4}$$

without the spatial temperature gradients, we have

$$\frac{\partial \sigma}{\partial t} - \beta(T)\frac{\partial}{\partial x}\left(\frac{\partial \sigma}{\partial x} - \alpha(j)\right) = 0 \tag{5}$$

Some example values for β under different temperatures are calculated using the data found in [2], [4], and shown in Table I.

Temperature(K)	350	370	400	420	440	450
$\beta(T)$	0.0029e-8	0.0080e-8	0.0302e-8	0.0659e-8	0.1336e-8	0.1857e-8

TABLE I

SOME VALUES OF β AT DIFFERENT TEMPERATURE. INTERCONNECT LENGTH OF 100μM IS USED.

Figure 1 illustrates the numerical solutions to Eq.(5) by showing the maximum stress build-up along the interconnect for several different values of $\beta(T)$, with the same initial and boundary conditions. Steady-state stress is determined by the term $q^*\rho j\,L/\Omega$. As seen in Figure 1, when plotted as a function of transformed time, solutions for different β values follow the same track before reaching their steady states. Also shown in Figure 1 are an example threshold stress $\sigma_{th} = 10^8$ Pa (which is within the range of typical values for critical stress failures) and the corresponding transformed time $\varphi_{th} = \beta(T)E(t_f)$. It follows that $E(t_f) = \varphi_{th}/\beta(T)$ is the true time to failure, and $1/E(t_f) = \beta(T)/\varphi_{th}$ is the failure rate. Notice that φ_{th} is constant for all of the interconnects that will eventually fail. In some cases, if $j{\cdot}L$ is less than a critical value $(j{\cdot}L)_c$, the steady-state stress build-up is below σ_{th}, and the interconnect will never fail [3].

If both current density and temperature change over the time, Equation (2) can be still used to model the EM stress build-up in the interconnect line. Clement [2] found that in the case where temperature is kept constant, the average current density should be used in Black's equation. In the case where temperature varies over time, we can assume that the average current density is used (i.e. α is a constant), while the temperature is a function of time (i.e. $\beta(t)$).

Since temperature may be an arbitrary function of time, it is impossible to directly solve the stress build-up equation. Instead, we solve the problem indirectly based on the following theorem.

Theorem 1: Consider the stress build-up equation (5) with constant values for β and α. Let $\sigma_1(x, t)$ be the solution for the equation with $\beta = \beta_1$ under certain initial and boundary conditions, $\sigma_2(x, t)$ be the solution with $\beta = \beta_2$ for the same initial and boundary conditions. If the solutions for Equation (5) are unique for those initial and boundary conditions, we have

$$\sigma_2(x, t) = \sigma_1(x, \left(\frac{\beta_2}{\beta_1}\right) t) \tag{6}$$

This theorem can be easily proved by plugging in Equation (5) with $\sigma_2(x, t)$ (Equation (6)) and β_2 respectively, and checking the satisfiability for the initial and boundary conditions.

Theorem (1) tells us that the stress build-up processes in the interconnect are independent of the value of β in Equation (5). The value of β only determines the build-up speed of the process. For example, at time $\left(\frac{\beta_2}{\beta_1}\right) t$, stress build-up in process 1 (i.e. $\beta = \beta_1$) *sees* the stress build-up in process 2 (i.e. $\beta = \beta_2$) at time t. If we are only concerned with the stress build-up at a fixed point on the line (e.g., at one end of the line, which always has the maximum stress at any time), we have the following corollary.

Corollary 1.1: For all stress build-up processes on the same metal line specified by Equation (5), with the only difference in the value of β, the stress build-up at a fixed point is only decided by the product of time and the value of β:

$$\sigma = f(\beta_i t_i) \tag{7}$$

where f is a function independent of temperature.

The corollary can be proven using Equation (6) with x fixed. Let σ_i denote the stress build-up at a specific point of the interconnect for the process with β_i. Corollary (1.1) shows that $\sigma_1(t_1) = \sigma_2(t_2) = \sigma_3(t_3) = \ldots$, as long as $\beta_1 t_1 = \beta_2 t_2 = \beta_3 t_3 = \ldots$.
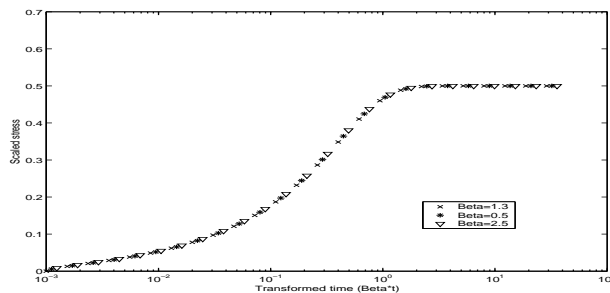

Fig. 2. EM stress build-up at one end of the interconnect for different β values as a function of βt.

Figure 2 illustrates Equation (7) by showing the stress build-up at one end of the metal line for several different values of β, with the same initial and boundary conditions. As expected from Equation (7), the stress is a unique function of βt, in spite of the different values of β.

Consider that temperature varies over time. We can divide time into segments, such that temperature is constant within each time segment. In other words, β in Equation (5) is a segment-wise function, described as:

$$\beta(t) = \begin{cases} \beta_1, & t \in [0, \Delta t_1] \\ \beta_2, & t \in (\Delta t_1, \Delta t_1 + \Delta t_2] \\ \cdots \\ \beta_i, & t \in \left(\sum_{k=1}^{i-1} \Delta t_k, \sum_{k=1}^{i} \Delta t_k\right] \\ \cdots \end{cases}$$

Since we are only concerned with the maximum stress build-up, let $\sigma_i(t)$ denote the distinct stress build-up at the end of a metal line with a constant value of β_i, and let $\sigma(t)$ be the stress build-up with the time-varying function $\beta(t)$. During the first time segment, the stress builds up the same as process $\sigma_1(t)$. At the end of this time segment, which is denoted by t_1 (i.e., $t_1 = \Delta t_1$), we have $\sigma(t_1) = \sigma_1(t_1) = f(\beta_1 \Delta t_1) = \sigma_2(\frac{\beta_1}{\beta_2} \Delta t_1)$, from Corollary (1.1). Also at time t_1, according to Theorem (1), the stress distribution along the metal line is the same as that for the process $\sigma_2(x, t)$ at time $\left(\frac{\beta_1}{\beta_2}\right) \Delta t_1$. Since during the second time segment, $\beta(t) = \beta_2$, the stress build-up (i.e.$(\Delta t_1, \Delta t_1 + \Delta t_2]$) is *exactly* the same as that in process $\sigma_2(t)$ during the time interval $\left(\left(\frac{\beta_1}{\beta_2}\right) \Delta t_1, \left(\frac{\beta_1}{\beta_2}\right) \Delta t_1 + \Delta t_2\right]$. So at the end of the second time segment (i.e. $t = t_2 = \Delta t_1 + \Delta t_2$), we have

$\sigma(t_2) = \sigma_2\left(\left(\frac{\beta_1}{\beta_2}\right) \Delta t_1 + \Delta t_2\right) = f\left(\beta_2 \left(\frac{\beta_1}{\beta_2} \Delta t_1 + \Delta t_2\right)\right) = f(\beta_1 \Delta t_1 + \beta_2 \Delta t_2)$, again due to Equation (7). Similar analysis can be applied to other time segments. It follows then that at the end the ith time segment, the stress is

specified as:

$$\sigma(t_i) = f\left(\sum_{k=1}^{i} \beta_k \Delta t_k\right) \tag{8}$$
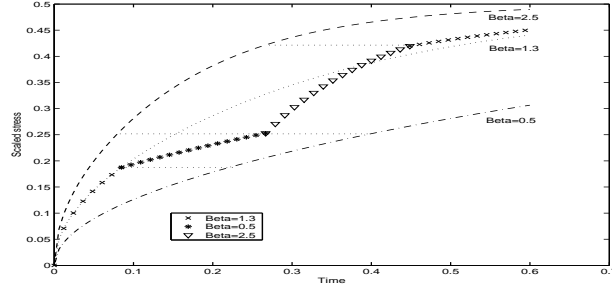
where f is defined in Corollary (1.1).

Fig. 3. EM stress build-up at one end of the interconnect with time-varying temperature. Also shown are the build-ups with invariant temperatures.

As an example, a numerical solution for Equation (5) with three different temperatures (i.e. $\beta_1 = 1.3, \beta_2 = 0.5, \beta_3 = 2.5$) over time is plotted in Figure 3. Also shown in the figure are the build-ups with time-invariant temperatures. As clearly indicated in the figure, the stress build-up process in each time segment is a shifted version of the corresponding process with time-invariant temperature. This figure is consistent with the above analysis.

As $\Delta t_i \to dt$, $\beta_i \to \beta(T(t))$, we obtain the integral version for the stress build-up function:

$$\sigma(t) = f\left(\int_0^t \beta(T(t))dt\right) \tag{9}$$

If we assume that the stress build-up reaches a certain threshold (σ_{th}) when an EM failure occurs, we have:

$$\int_0^{t_{failure}} \beta(T(t))dt = \varphi_{th} \tag{10}$$

where φ_{th} is decided by the critical stress (i.e. $\varphi_{th} = f^{-1}(\sigma_{th})$). If an average value of $\beta(t)$ exists, we obtain a closed form for the time to failure:

$$t_{failure} = \frac{\varphi_{th}}{E(\beta(T(t)))} \tag{11}$$

where $E(\beta(t))$ is the expected value for $\beta(t)$, and $\beta(t)$ is a function of temperature having the form:

$$\beta(T(t)) = A\left(\frac{exp\left(-\frac{Q}{kT(t)}\right)}{kT(t)}\right) \tag{12}$$

where A is a constant.

A key implication of this new model is the *increased flexibility for dynamic thermal management*. Existing techniques typically specify a fixed temperature threshold based on a defined lifetime goal, and the hardware must ensure that the operating temperature never exceeds that threshold. Our analysis shows that this limitation is overly conservative. Instead, Eq.(10) reveals that interconnect lifetime can be modeled as a resource that is consumed during circuit operation at a rate of $\beta(T)$. Therefore, it is safe to overshoot the fixed temperature threshold for limited periods of time in order to obtain higher circuit performance, and then later compensate for the excess lifetime "consumption" with lower temperatures to ensure the lifetime requirements are still satisfied. Other dynamic thermal management techniques that exploit the flexibility provided by this new model will be explored as part of future work.

In addition, it can be shown that $\beta(T(t))$ is a convex function with respect to temperature within the normal operating temperature range. By applying Jensen's inequality for convex functions, we have

$$E(\beta(T(t))) \geq \beta(E(T(t)))$$

which, according to Eq.(11), leads to another key observation: *a constant temperature T will always yield a longer expected lifetime than a time-varying temperature with an average of T*. Damping temporal variations can therefore permit a higher operating temperature and performance while maintaining the specified expected lifetime.
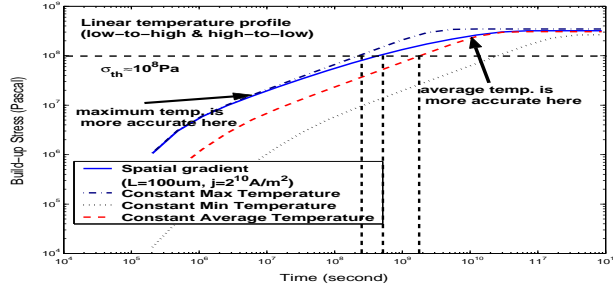
Fig. 4. Solutions for a linear spatial temperature gradient along a wire, together with solutions to constant max, min and average temperatures.
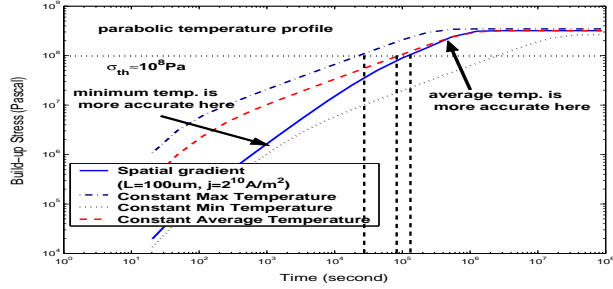

Fig. 5. Solutions for a parabolic spatial temperature gradient along a wire, together with solutions to constant max, min and average temperatures.

B. Spatial temperature gradients

The other aspect of temperature-aware interconnect EM analysis is the effect of temperature variation over the length of a wire, that is, interconnect temperature as a function of position, $T(x)$. This is a concern because atom diffusivity, which is exponentially dependent on temperature as shown in Eq.(3), varies along the wire with temperature. This leads to a different stress build-up distribution than that for a spatially constant wire temperature.

To illustrate the importance of spatial gradients, we consider several simple temperature profiles along a wire: 1) current flows from a hot to a cold region of the design with a linear temperature profile; 2) the reverse, with current flowing from cold to hot with a liner temperature profile; and 3) a temperature profile that is parabolic, with the lowest temperatures at the ends and maximum at the center of the wire, as might be found if the wire is within a circuit area of uniform power distribution that is surrounded by a colder region.

By numerically solving Eq.(2) for each of the three temperature profiles, we can plot the maximum stress build-up on the wire over time, as shown in Figures 4–5. These plots also include the stress build-up curves corresponding to three different spatially constant temperatures: maximum temperature $T_{max} = 150°C$, minimum temperature $T_{min} = 60°C$ and the average wire temperature for the spatial gradient profiles above). The critical stress for the wire to fail is again set to 10^8Pa.

It is clear from Figures 4–5 that the actual stress build-up of a wire with spatial temperature gradients cannot be accurately modeled using a spatially constant temperature. With the linear temperature profiles in Figure 4, the actual stress build-up is bounded above by that of the maximum temperature and below by that of the average temperature. Interestingly, when the stress build-up is small (e.g., less than 10^7Pa), the spatial gradient EM can be well approximated by that of the constant maximum temperature, but when the stress build-up is close to saturation, the spatial gradient plot is similar to that of the constant average temperature.

Although the high-to-low and low-to-high temperature profiles have the same maximal values of stress build-up over time, the direction of the stress in these two cases are opposite. A wire with a low-to-high temperature gradient in the direction of the current is more prone to an "open" failure, because the diffusion of metal atoms (i.e., tensile stress) is more severe at the high-temperature end of the wire. Conversely, "short" failures are more common for wires with high-to-low temperature gradients in the direction of the current, because accumulation of metal atoms (i.e., compressive stress) is more severe at the high-temperature end of the interconnect.

Of even greater interest is the parabolic gradient in Figure 5, where the stress build-up is bounded above by the *average* case, even though some parts of the wire operate at temperatures higher than that average. Detailed analysis reveals that the symmetric temperature distribution along the wire has a damping effect on atom migration, reducing the diffusion speed incurred by the constant maximum temperature.

In the above analysis, we assume that the spatial profile of the temperature is unchanged over time. Future work includes developing an interconnect reliability model that combines both temporal and spatial gradients—in other words,
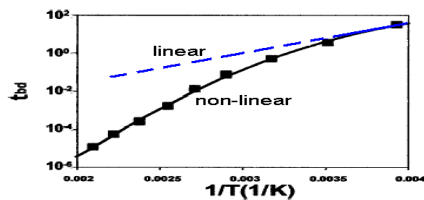
Fig. 6. Arrhenius plot for thin oxide breakdown showing non-linear trend with greater slope at high temperature, which means the breakdown process exacerbates at higher temperatures for thinner gate oxide. (adapted from [1])

the spatial gradient itself exhibits variations over time. So far we are proposing a conservative approach to solve this problem by finding the worst spatial gradient and applying our analysis to that.

III. GATE-OXIDE RELIABILITY ANALYSIS WITH TEMPERATURE GRADIENTS

The correct physical model for gate-oxide breakdown is far from being settled. Here we merely wish to observe that the same approach that we described in the previous section will also improve the precision of modeling gate-oxide breakdown. We base our analysis here on work by Cheung [1], who argues that thin oxide breakdown is due to the accumulation of neutral defects. The involved reactions are

$$V + h^+ \underset{k_3}{\overset{k_2}{\rightleftharpoons}} P \overset{k_4}{\underset{e^-}{\longrightarrow}} N$$
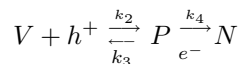
where V is a precursor site, P is a intermediate site, and N is a neutral defect. He proposes a model for the calculation of neutral defect concentration, which can well explain the fast temperature acceleration in ultra-thin oxide-breakdown and the nonlinear Arrhenius behavior. The model is

$$[N(t)] = \frac{[V]_0 k_2 J_h k_4 J_e}{(k_3 + k_4 J_e)^2}(1 - e^{-(k_3 + k_4 J_e)t}) \tag{13}$$

where $[N(t)]$ is the time-dependent neutral defect concentration which has a threshold value $[N]_{th}$. If $[N(t)] > [N]_{th}$, the oxide is considered broken-down; $[V]_0$ is the initial concentration of the precursor sites which can be converted to neutral defect sites by first capturing holes/protons followed by capturing electrons. J_h and J_e are the hole and electron current density through the gate oxide. k_2, k_3 and k_4 are the rates for the above microscopic reactions, and they all have the Arrhenius form of temperature dependency (B_2, B_3 and B_4 are constants.)

$$k_{2,3,4} = B_{2,3,4} exp\left(\frac{-Q_{2,3,4}}{kT}\right)$$

Thus, if we define

$$\gamma(T) = k_3 + k_4 J_e = B_3 exp\left(\frac{-Q_3}{kT}\right) + J_e B_4 exp\left(\frac{-Q_4}{kT}\right) \tag{14}$$

and consider the term $\alpha(T) = ([V]_0 k_2 J_h k_4 J_e)/(k_3 + k_4 J_e)^2$ in Eq.(13) as a constant because $\alpha(T)$ is weakly dependent on temperature compared to the term $e^{-\gamma(T)t}$, which is super-exponential, we have

$$[N(t)] \approx \alpha(1 - e^{-\gamma(T)t}) \tag{15}$$

We find that we can use exactly the same method we used in Section 2(A) when analyzing the effect of temporal temperature gradients on interconnect reliability: define $\psi_{th} = \gamma(T)t_{bd}$ as the transformed time at which the defect concentration reaches $[N]_{th}$. Therefore, $t_{bd} = \psi_{th}/\gamma(T)$ is the time-to-breakdown.

If temperature varies over time (i.e., there is temporal temperature gradient $T(t)$ at a particular device) we have

$$\psi_{th} = \int_0^{t_{bd}} \gamma(T(t))dt$$

Thus, $\gamma(T(t))$ is the rate at which the gate-oxide's lifetime is consumed. Similar to Section 2(A), we can argue that analysis based on a uniform worst-case temperature is overly conservative. We have only considered temporal temperature gradients in this section because CMOS devices are so tiny that spatial temperature gradients do not apply to individual devices for gate-oxide reliability analysis.

IV. Temperature-Aware Design Using Temperature Gradients

A. Analytic analysis

Our proposed dynamic reliability model reveals that lifetime reliability can be modeled as a resource to be consumed over time at a temperature dependent consumption rate. This provides flexibility during runtime management. For example, our model gives the direct relation between reliability and performance. Users can trade off reliability with performance. During operation, if the processor runs below the target temperature, lifetime is consumed at a rate less than expected. Our model infers that the processor is allowed to operate at a higher temperature later for certain interval, thus maximizing the performance while satisfying the reliability requirement. In the following, we use a simple analytic model to demonstrate how our reliability model can be used to guide runtime management.

Processor lifetime consumption rate can be expressed as $r(t) = A\frac{j(t)}{k_b T(t)}e^{-\frac{E_a}{k_b T(t)}}$, where A is a constant, k_b is Boltzmann constant, E_a is the activation energy, $T(t)$ is temperature and $j(t)$ is current density (we assume a void-induced EM failure, which is commonly seen in modern ICs, and the exponent for current density is 1). Since j is proportional to the product of voltage and frequency, and assuming that clock frequency is proportional to voltage, we can rewrite the above equation as: $r(t) = A\frac{f^2}{k_b T(t)}e^{-\frac{E_a}{k_b T(t)}}$. Using ITRS data, f=3.0GHz, voltage = 1.2V, temperature = 105C and a simple relation $T \approx K_v V^3 + T_0 \approx K_f f^3 + T_0$ (with $T_0 = 318K$ as the ambient temperature, including the effect of static power), we obtain $K_f = 2.22K/GHz^3$. And the reliability consumption rate can be re-written as: $r(f) = A\frac{f^2}{k_b(K_f f^3 + T_0)}e^{-\frac{E_a}{k_b(K_f f^3 + T_0)}}$. Therefore, we can trade-off reliability r with performance f by:

$$r(f_1) = A\frac{f_1^2}{k_b(K_f f_1^3 + T_0)}e^{-\frac{E_a}{k_b(K_f f_1^3 + T_0)}} = \gamma r(f_2) = \gamma A\frac{f_2^2}{k_b(K_f f_2^3 + T_0)}e^{-\frac{E_a}{k_b(K_f f_2^3 + T_0)}}$$

where $r(f_1)$ is the nominal consumption rate at the nominal frequency f_1, and γ is the trade-off factor for lifetime (e.g. $\gamma = 1/2$ means shortening the lifetime by a half). Table II presents some examples of trade-off between lifetime and performance. It clearly indicates that, without temperature variations, a small improvement in performance can only be achieved at great sacrifice of lifetime.

γ	0.8	0.5	0.3	0.1
f_2(GHz)	3.05	3.2	3.36	3.65

TABLE II

Trade-off between lifetime and performance (nominal clock frequency is 3.0GHz)

Due to temperature variations during system operations, a processor will be operated under the nominal temperature for some time, which brings the opportunities for higher performance by increasing the clock frequency (thus at higher temperature) later. For instance, if the processor is operated at a temperature 10^oC below 105^oC (nominal temperature) for a period of time $t1$ (cool time), we can estimate how much performance gain we will have in terms of the time interval allowable to run at a higher frequency. Let $t2$ denote the allowable time interval to run at a higher temperature (hot time), we have:

$$(A\frac{1}{k_b T_1}e^{-\frac{E}{k_b T_1}} - A\frac{1}{k_b(T_1 - 10)}e^{-\frac{E}{k_b(T_1 - 10)}})t_1 = (A\frac{1}{k_b T_2}e^{-\frac{E}{k_b T_2}} - A\frac{1}{k_b T_1}e^{-\frac{E}{k_b T_1}})t_2$$

where T_1 is the nominal temperature (i.e. 105^oC), T_2 is the allowable higher temperature during time interval t_2.

Table III lists some examples for over-clocking opportunities. For example, assuming that the chip nominal interconnect temperature at the nominal clock frequency (e.g. 3.0 GHz) is 105^oC, if the temperature drops by 10^oC due to workload variations for a time interval t_1, our calculation indicates that we can over-clock the chip to 3.16GHz for an interval of $t_2 = 0.738t_1$, without violating the reliability specification.

B. A grid-like compact thermal model

In the following, we present an interconnect reliability case study showing the benefits of temperature-aware reliability analysis. The effects of temporal and spatial temperature gradients on EM failure are investigated for a simulated 0.13μm microprocessor. Interconnect temperatures are obtained from a new, validated compact thermal model we have developed that divides the chip area into a fine grid of equal-area cells. We briefly introduce the thermal model and its utility to the VLSI design community, and then present and discuss the simulation results. In this paper, we do not perform a case study for gate-oxide reliability analysis due to the fact that the appropriate values of k_3 and k_4 were not discovered in the literature. Such a study will be part of future work, but based on the similarity of the analysis, we

Running 10^oC below nominal temperature for t_1				
Temperature above 105°C (threshold)	5	10	15	20
Clock frequency (GHz)	3.0821	3.1592	3.2327	3.303
Ratio of hot time over cool time (t2/t1)	1.4783	0.738	0.452	0.31

Running 15^oC below nominal temperature for t_1				
Temperature above 105°C (threshold)	10	15	20	25
Clock frequency (GHz)	3.1592	3.2327	3.303	3.3705
Ratio of hot time over cool time (t2/t1)	0.9071	0.5447	0.3677	0.2646

Running 20^oC below nominal temperature for t_1				
Temperature above 105°C (threshold)	15	20	25	30
Clock frequency (GHz)	3.2327	3.303	3.3705	3.4353
Ratio of hot time over cool time (t2/t1)	0.6604	0.4458	0.3209	0.2404

TABLE III

OVER-CLOCK OPPORTUNITIES BY DYNAMIC RUNTIME RELIABILITY MANAGEMENT FOR THE CASE WHEN TEMPERATURE IS DROPPED BELOW THE NOMINAL TEMPERATURE DUE TO WORKLOAD VARIATIONS. (NOMINAL CLOCK FREQUENCY IS 3.0GHz)

expect that accounting for temporal temperature variation in gate-oxide reliability would yield similar benefits as for interconnect EM.

The accuracy of reliability analysis using the models presented here depends greatly on accurate and detailed temperature estimations. A useful model must be able to simulate transient, not just steady-state, temperatures. It should also be parameterized so that a correct model is generated regardless of the materials, layout, or thermal package. Finally, different levels of granularity should be supported, so that thermal analysis can be a part of the design process from early architectural studies through different circuit-design stages, including detailed design, placement, and routing.

We have extended our prior HotSpot model [6], which meets the first two requirements but was only designed to model temperature at a microarchitecture level of granularity. Instead of the ad-hoc structure of that original model, we have developed a grid-based model that accommodates the required range of granularities. The grid-based model is also necessary for obtaining data for spatial temperature gradients. For example, the grid can be set to the granularity of individual fundamental circuit structures, standard cells, functional units, etc. The grid-based approach is therefore more general. Its adjustable grid size enables designers to perform temperature-related analysis at any level of granularity, including the microarchitecture level targeted by the original model. It is also useful to note that the new model can be easily extended to a multi-resolution grid with fine resolution in critical portions of the circuit and coarse resolution elsewhere.

Figure 7 shows the structure of the model. It consists of a grid of cells, with each cell consisting of thermal resistors and capacitors representing silicon, heat spreader, thermal interface material and heat sink. Validation shows that an essential component of the model (in addition to the detailed model of the die and the various components of the thermal package) is the thermal interface material between the layers of the thermal package. (After modeling the interface material, the spatial temperature gradient across the die increases to a range of 30 to 50 degrees, which is more accurate than the gradient of about 10 degrees predicted in [6].) This extended thermal model is also able to predict interconnect temperatures by adding layers of thermal resistors and capacitors representing interconnect metal layers together with inter-layer dielectrics and vias that are above each of the silicon grids. The numerical solution for the grid-based compact model follows the same algorithm used in [6].

Another important improvement over our prior work is physical validation of the thermal model, which we have performed by comparing the estimated silicon surface temperatures from the model against those obtained with a commercial thermal test chip [7]. The test chip has a 9x9 grid of power dissipating resistors, which can be turned on or off individually, and a corresponding 9x9 grid of temperature sensors; and can measure both steady-state and transient temperatures for each of the power dissipators. We build the same 9x9 grid-like chip structure in our thermal model, turn on specific sets of power dissipators in the test chip, and assign exactly the same power values at the same locations in the grid-based thermal model. Figure 8 (I) compares steady-state thermal plots for one particular pattern; and Figure 8 (II) compares transient temperature measurements. Power density in this experiment is 0.5W/mm^2 in the heat-dissipating area. As can be seen, our compact thermal model is quite accurate, with the worst case error for steady-state and transient temperatures less than 5% and 7%, respectively. Similar results were obtained for other power-dissipation patterns.
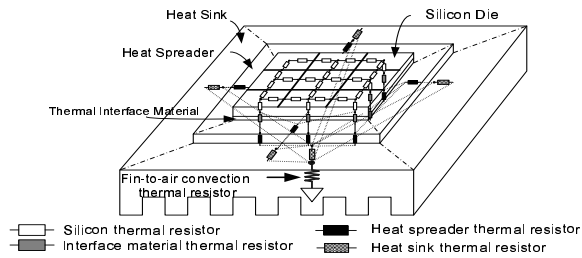
Fig. 7. Example structure of the grid-like compact thermal model with 3x3 grids, thermal interface material, heat spreader and heat sink. Thermal capacitors and heat sources are not drawn for clarity.
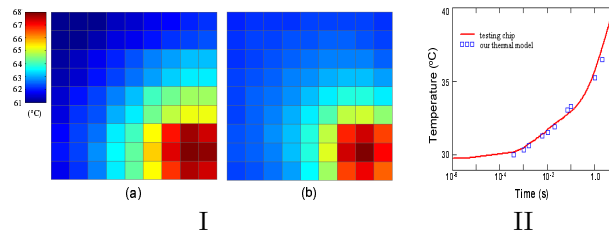

Fig. 8. (I)— Steady-state validation of the compact thermal model: (a) Testing chip measurements (b) Results from the model with errors less than 5%. (II)— Transient validation of the compact thermal model. Percentage error is less than 7%. (Transient temperature response of one power dissipater is shown here.)

C. A case study: Impact of temperature gradients for typical program behavior

To demonstrate the benefits of accounting for temporal and spatial temperature gradients, we present a case study using temperature values obtained from simulating a microprocessor with characteristics similar to a 0.13μm Alpha 21364. Using the described grid-based compact thermal model, we can obtain the steady-state and transient temperature responses of the devices and interconnect. For example, the transient thermal behavior of interconnect above the floating point register is shown in Figure 9 for the SPEC2000 benchmark program *applu*, revealing obvious temporal temperature gradients. The thermal-package characteristics we used in simulating applu were derived so that the constant temperature value that, according to Section 2A, yields the same expected lifetime as the pattern in Figure 9 is $110°$C (a common limit). This temperature has also been plotted in Figure 9 as a straight line.

These results illustrate the potential benefits of accounting for temporal variation. If the lifetime budget is used to dictate only some fixed worst-case temperature (e.g., $110°$C), then a more expensive cooling solution is required to bring applu's actual behavior within specification while achieving the same performance. The alternative is that the voltage and clock speed must be reduced, or a dynamic thermal management technique must be engaged to reduce processor activity and enforce the $110°$ limit whenever the operating temperature exceeds the threshold. Using microarchitecture simulation techniques described in [6], we estimate that selecting a lower design point for voltage and frequency would require a 13% reduction in clock frequency; and that dynamic thermal management would reduce performance by about 10% using dynamic voltage scaling and 50% using fine-grained fetch gating. If temporal temperature variation is taken into account, none of these costly solutions are needed!

Now consider spatial temperature gradients along the interconnect. From the compact thermal model, for a typical point in time we find that the temperature of the interconnect above the floating point register is about $110°$C, while the temperature for the interconnect above the adjacent floating point queue is about $73°$C, a $37°$C gradient for interconnect
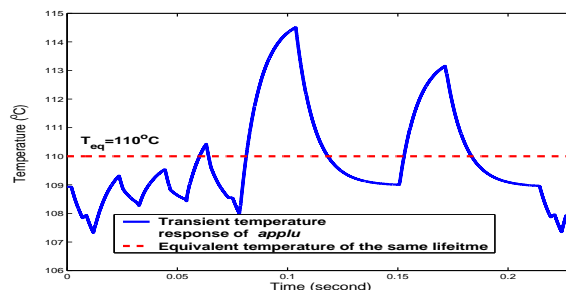

Fig. 9. Transient temperature response of floating point register interconnects. Constant operating temperature for the same interconnect lifetime is also shown.
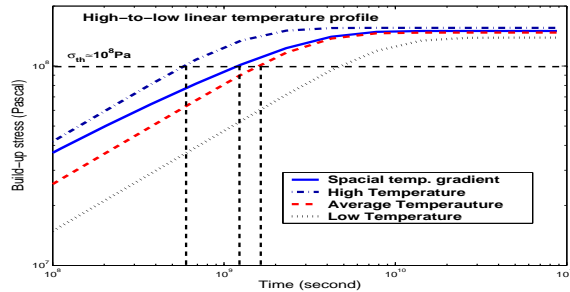
Fig. 10. Lifetime prediction comparison for a wire with and without a spatial temperature gradient.

from the floating point register to the floating point queue. From our temperature-aware analysis shown in Figure 10, which plots the stress build-up over time for this particular spatial-gradient pattern, we can see that assuming a uniform average temperature along the interconnect overestimates expected lifetime by 30% compared to the actual case, and using a uniform maximum temperature underestimates expected lifetime by 80%. This means that the typical approach of worst-case analysis will yield drastically lower temperature specifications than necessary, which again would require an unnecessarily expensive cooling solution or unnecessary performance sacrifices.

V. Conclusions

In this paper, we have presented a new approach to interconnect and device gate-oxide reliability analysis that accounts for temporal and spatial variations in temperature. We have also developed *and validated* a dynamic compact thermal model for simulating, at various levels of detail, the time-dependent evolution of on-chip temperatures across an IC. Reliability analysis using temporal and spatial gradient values obtained from a real application on a simulated processor show the importance of accounting for temperature gradients. Worst-case analysis can drastically underestimate expected lifetime, requiring either unnecessarily aggressive and costly cooling solutions or else reductions in power dissipation that incur unnecessary sacrifices in IC performance. We propose that, instead of designing for a maximum tolerated temperature based on a worst-case analysis, expected lifetime should be viewed as a resource that is consumed over time at a temperature-dependent rate. This dynamic, reliability-driven approach to managing operating temperature fits particularly well with the recent advent of dynamic thermal management techniques; and this paper shows that lifetime requirements are the proper objective function rather than fixed temperature thresholds. This modeling approach will be extended to other failure mechanisms and integrated with techniques for real-time thermal management.

Acknowledgments

References

[1] K. P. Cheung. Temperature effect on ultrathin sio$_2$ time-dependent dielectric breakdown. *Applied Physics Letter*, 2003.
[2] J. Clement. Reliability analysis for encapsulated interconnect lines under dc and pulsed dc current using a continuum electromigration transport model. *J. Appl. Phys.*, 82(12):5991, 1997.
[3] J. Clement. Electromigration modeling for integrated circuit interconnect reliability analysis. *IEEE Transactions on Device and Materials Reliability*, 1(1):33–42, March 2001.
[4] M. A. Korhonen and P. Bøgesen. Stress evolution due to electromigration in confined metal lines. *J. Appl. Phys.*, 73(8):3790, April 1993.
[5] C. J. M. Lasance. Thermally driven reliability issues in microelectronic systems: status-quo and challenges. *Microelectronics and Reliability*, 2003.
[6] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankanarayanan, and D. Tarjan. Temperature-ware microarchitecture. In *Proceedings of the 30th International Symposium on Computer Architecture*, pages 2–13, San Diego, CA, 2003.
[7] V. Székely, M. R. C. Márta, G. Végh, Z. Benedek, and S. Török. A thermal benchmark chip: design and applications. *IEEE Trans. on Components, Packaging, and Manufacturing Technology–Part A*, 1998.