Analysis and Optimization of Thermal Issues in High-Performance VLSI

Kaustav Banerjee
Center for Integrated Systems
Stanford University, Stanford, CA 94305
kaustav@cis.stanford.edu

Massoud Pedram and Amir H. Ajami
Department of Electrical Engineering-Systems
University of Southern California, Los Angeles, CA 90089
{massoud, aajami}@zugros.usc.edu

Abstract

This paper provides an overview of various thermal issues in high-performance VLSI with especial attention to their implications for performance and reliability. More specifically, it examines the impact of thermal effects on both interconnect design and electromigration reliability and discusses their impact on the allowable current density limits. Furthermore, it also discusses how thermal and reliability constrained current density limits may conflict with those obtained through purely performance based criterion. Additionally, it is shown that chip level thermal effects can have a significant impact on large-scale circuit optimization techniques, including the clock-skew minimization scheme, and can influence other physical design problem formulations. Finally, high-current interconnect design rules for ESD and I/O circuits are also examined.

1 Introduction

As VLSI technology scales, thermal issues are becoming the dominant factor in determining performance, reliability and cost of high-performance ICs. Management of these issues is going to be one of the key factors in the development of next generation microprocessors, integrated networks, and other highly integrated systems.

1.1 Sources of Power Dissipation

For digital CMOS circuits there are four sources of power dissipation and the average power dissipation, P_{avg}, can be expressed as [1],

$$P_{avg} = P_{switching} + P_{short-circuit} + P_{leakage} + P_{static} \quad (1)$$

where $P_{switching}\left(=0.5\,\alpha C V_{dd}^2 f\right)$ is the switching component of power, C is the load capacitance, f is the clock frequency, and α is the node transition activity factor ($0 \le \alpha \le 1$). $P_{short-circuit}\left(=I_{sc}\cdot V_{dd}\right)$ is due to the direct path short-circuit current, I_{sc}, which arises when both the NMOS and PMOS transistors are simultaneously active, conducting current directly from supply to ground. $P_{leakage}\left(=I_{leakage}\cdot V_{dd}\right)$ is due to the leakage current, $I_{leakage}$, which can arise from reverse bias diode currents and sub-threshold effects. Finally, $P_{static}\left(=I_{static}\cdot V_{dd}\right)$, is due to the static current arising from circuits that have a constant source of current between the power supplies. These four components of power dissipation are all associated with the devices. Also, the most dominating component is the one due to switching, although the leakage component is becoming increasingly significant for deep sub micron technologies. It should be noted that the capacitance C in the switching component of power dissipation is predominantly due to the interconnect capacitances. Hence, it is the interconnects that are mainly responsible for the total chip power dissipation.

ISPD'01, April 1-4, 2001, Sonoma, California, USA.

1.2 Chip Temperature Estimation

The average temperature of the chip, T_{chip}, can be estimated using the following equation,

$$T_{chip} = T_0 + R_n\left(\frac{P}{A}\right) \quad (2)$$

where T_0 is the ambient temperature, P is the average power dissipation and A is the chip area. Here R_n represents the substrate (Si) layer plus the package thermal resistance. Using (2) R_n was found to be 4.75 cm^2 °C/W, based on the operating chip temperature (T_{chip}=120 °C) for the present technology node (180 nm) and T_0=25 ^{0}C. Assuming the same value for R_n, the die temperatures at other technology nodes can be estimated using (2).

In addition to the four components mentioned above some power dissipation also results from Joule heating (or self-heating) caused by the flow of current in the interconnect network. Although interconnect Joule heating constitutes only a small fraction of the total power dissipation in the chip, the temperature rise in the interconnects due to Joule heating can be significant. This is due to the fact that interconnects are located away from the Silicon substrate and the heat sink by several layers of insulating materials which have lower thermal conductivities than that of Silicon. In fact, full chip thermal analysis using Finite Element simulations has recently shown that the maximum temperature in the chip increases rapidly with scaling due to increased Joule heating of the interconnects [3] despite the chip power density (power per unit area) remaining nearly constant over a wide range of technology nodes as per the ITRS [2]. The maximum temperature occurs at the top of the chip where the global interconnects are located. Hence, thermal effects in interconnects require careful consideration.

1.3 Thermal Effects in Interconnects

Thermal effects in interconnects are an inseparable aspect of electrical power distribution and signal transmission through the interconnects in VLSI circuits due to self-heating (or Joule heating) caused by the flow of the current [4]. Thermal effects impact interconnect design and reliability in the following ways. First, they limit the maximum allowable RMS current density, $j_{RMS-max}$ (since the RMS value of the current density is responsible for heat generation), in the interconnects, in order to limit the temperature increase. Second, interconnect lifetime (reliability), which is limited by electromigration (EM), has an exponential dependence on the inverse metal temperature [5]. Hence, the temperature rise in metal interconnects due to the self-heating phenomenon can also limit the maximum allowed average current density, $j_{avg-max}$, since EM capability is dependent on the average current density [6]. Third, thermally induced open circuit metal failure under short-duration high peak currents, including electrostatic discharge (ESD), is also a reliability concern, [7] and can introduce latent EM damage [8] that has important reliability implications.

Additionally, low dielectric constant (low-k) materials are being introduced as an alternative intra- and inter- level insulator to reduce interconnect capacitance (and therefore delay) and cross-talk noise to enhance circuit performance [9],[10],[11]. These materials can further exacerbate thermal effects owing to their lower thermal conductivity in comparison to that of silicon dioxide [12]. Furthermore, it is important to understand how thermal and reliability constraints may conflict with the

performance optimization steps employed at the circuit level. Hence, thorough analysis of thermal effects in DSM interconnects is necessary to comprehend their full impact on circuit design, accurately model their reliability, and provide thermally safe design guidelines for various technologies.

1.4 Trends in Interconnect Scaling and Implications for Thermal Effects

As VLSI circuits continue to be scaled aggressively, a rapid increase in functional density and chip size is observed [2]. This has resulted in increasing number of interconnect levels and reduction in interconnect pitch in order to realize all the inter-device and inter-block communications. Interconnect levels are expected to increase further in the near future, from 6 levels at the 250 nm node to 9 levels at the 50 nm node. This increase in the number of interconnect levels causes the upper most interconnect layers to move further away from the Si substrate, making heat dissipation more difficult. Additionally, decreasing interconnect pitch will cause increased thermal coupling. Furthermore, the critical dimensions of contacts and vias are also decreasing with scaling, resulting in higher current densities in these structures. Compounded with the introduction of low-k dielectrics, it is envisioned that thermal effects in interconnects could become another serious design constraint.

Thus the various trends in technology scaling that cause increased thermal effects in interconnects can be summarized as:

- Increasing current density
- Increasing the number of interconnect levels
- Introduction of low-k dielectric materials
- Increased thermal coupling

1.5 Implications of Thermal Effects on Reliability and Performance

In current VLSI interconnect designs, current density design rules are typically based on EM lifetimes [6]. However, the actual current densities are determined by the interconnect parameters (resistance and capacitance per unit length and the length of the interconnect) and the strength of the buffer that is driving the interconnect length. In a typical design, long interconnects, which can potentially have large current densities, are usually split into buffered segments to improve performance (signal skew and delay) [13]. The signal line length and buffer sizes are optimized to give maximum performance for a given technology. Signal lines longer than the optimum length would increase interconnect delay while buffer sizes larger than the optimum would result in an increase in buffer delay. Moreover, increasing buffer size would increase the current density of the signal line connected to the output and may also cause excessive power dissipation. Hence, it is important to quantify whether the associated current densities in signal lines optimized for maximum performance also meet the EM design limits.

As mentioned earlier, in DSM technologies, low-k materials invariably have poor thermal properties, and therefore the use of such materials can significantly impact the EM design limits. On the other hand, the use of Copper, which has a lower resistivity, alleviates the problem to some extent. It is therefore important to quantify whether EM reliability or performance is the dominant factor in determining the optimal signal line length in various low-k/Cu based interconnect systems.

Section 2 examines the interconnect thermal profile analysis by using the heat diffusion equation. In Section 3, the effect of self-heating on EM is briefly discussed followed by an analysis of the self-consistent design approach. The self-consistent approach is then used to quantify the impact of new interconnect materials (Cu and low-k dielectrics) on allowed current density limits. Section 4 introduces the methodology for computing the current densities from performance considerations only, and then provides a direct comparison between the reliability and performance based current density design limits [14], [15]. In Section 5, the effects of temperature on performance metrics (i.e. delay and clock skew) are studied. Using a recently developed temperature-dependent delay model, design rules for maintaining zero-skew clock trees in high performance designs are provided. Finally, in Section 6, interconnect design rules necessary for electrostatic discharge (ESD) and I/O circuits are discussed.

2 Interconnect Thermal Profile

2.1 Non-uniform Chip Thermal Profile

The main sources of temperature generation in the chip are the switching activities of the cells over the substrate. In a high performance design the substrate temperature can reach up to 120 °C, and Joule heating can contribute further to the overall temperature of an interconnect [3].

Due to the presence of many heat generation sources in the substrate and the complicated boundary conditions (because of the convective nature of the heat transfer between the bottom side of the heat sink and the ambient), finding an analytical solution for the heat diffusion equation is non-trivial. As a result, much of the research work has focused on obtaining a solution by using numerical techniques, most notably the fast thermal analysis (FTA) method [16]. However, it must be mentioned that the accuracy of this kind of analysis depends on how accurately the power consumption of each cell (or macro-cell) over the substrate can be determined. In general, the thermal profile over the surface of the substrate depends on the power consumption of the cells and the distances between them.

2.2 Analytical Model for Interconnect Temperature Profile

Using appropriate boundary conditions, heat flow in interconnect can be obtained by solving the heat diffusion equation in the *3-D* space. In the steady state, by assuming that the four sidewalls and the top surface of the chip are thermally isolated (these are generally valid assumptions), the heat diffusion equation can be reduced to a *1-D* form as follows:

$$\frac{d^2T}{dx^2} = -\frac{Q}{k_m} \qquad (3)$$

where Q is the volumetric heat generation rate inside the interconnect (W/m^3) and k_m is the thermal conductivity of the interconnect material (W/m°C) which is assumed to be constant. Consider an interconnect with length L, width w and thickness t_m that passes over the substrate with an insulator of thickness t_{ins} and thermal conductivity k_{ins} separating the two. The interconnect is connected to the substrate by vias/contacts at its two ends. The volumetric heat generation in the interconnect is computed by determining the rate of power generation due to the RMS current and the rate of heat loss due to the heat transfer between the interconnect and the substrate through the insulator. As a result, the heat flow equation (3) in an interconnect can be restated as follows [17]:

$$\frac{d^2T_{line}(x)}{dx^2} = \lambda^2 T_{line}(x) - \lambda^2 T_{ref}(x) - \theta \qquad (4)$$

where λ and θ are constants given as follows:

$$\lambda^2 = \frac{1}{k_m}\left(\frac{k_{ins}}{t_m \cdot t_{ins}} - \frac{I_{rms}^{\ 2} \cdot \rho \cdot \beta}{w^2 t^2}\right) \qquad (5)$$

$$\theta = \frac{I_{rms}^{\ 2} \cdot \rho}{w^2 \cdot t_m^{\ 2} \cdot k_m} \qquad (6)$$

T_{line} is the interconnect temperature as a function of position along the length of the interconnect (which we will refer to as the interconnect thermal profile), T_{ref} is the underlying substrate temperature, ρ is the metal electrical resistivity at the reference temperature (0 °C), and β is the temperature coefficient of resistance in 1/°C (see (21)). In order to have a unique solution for (4), we need to provide two boundary conditions. Equation (4) shows the importance of the substrate temperature profile T_{ref} in determining the interconnect temperature. When considering short local wires, T_{ref} is usually assumed to be a constant. For long global interconnects, this may not always be a valid assumption since these lines span a large area of the substrate surface. Due to different switching

activities of the cells in the substrate, a non-uniform temperature gradient which is created by the so-called *hot spots* over the substrate, is inevitable. As a result, determining the substrate thermal profile is crucial to the thermal analysis of the interconnects.

From (5) and (6) it can be deduced that the thermal profile along an interconnect is strongly dependent on the thickness of the underlying insulator t_{ins}. It is obvious that for a given technology, interconnects assigned to higher metal layers are farther from the substrate and as a result, the thermal resistances between these interconnects and the substrate are larger. Therefore, the higher metal layers that carry the clock signal (and have higher current density than the lower metal layers [14]) experience higher temperatures in comparison to the lower metal layers (recall that the interconnect can only exchange energy with the substrate and the top side of the chip is assumed to be thermally isolated). Figure 1 shows the thermal profiles along the length of a 2000 µm long interconnect.

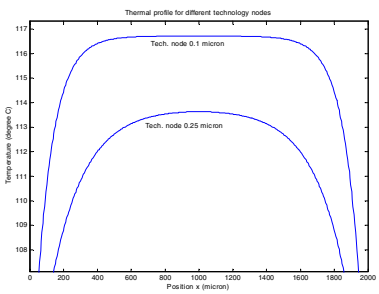

Figure 1: Thermal profile along the length of an interconnect with uniform substrate temperature using interconnect parameters for 0.1 and 0.25 µm technology nodes.

3 Interconnect Reliability

3.1 Influence of Self-Heating on EM

EM lifetime reliability of metal interconnects is modeled by the well known Black's equation [5] given by,

$$TTF = A\, j^{-n}\, exp\left(\frac{Q}{k_B T_m}\right) \tag{7}$$

where TTF is the time-to-fail (typically for 0.1% cumulative failure). A is a constant that is dependent on the geometry and microstructure of the interconnect, j is the DC or average current density. The exponent n is typically 2 under normal use conditions, Q is the activation energy for grain-boundary diffusion and equals ~ 0.7 eV for Al-Cu, k_B is the Boltzmann's constant, and T_m is the metal temperature. The typical goal is to achieve 10 year lifetime at 100 ^{0}C, for which (7) and accelerated testing data produce a design rule value for the acceptable current density, j_0, at the reference temperature T_{ref}. However, this design rule value does not comprehend self-heating.

The effect of self-heating can be analyzed from the following: By simplifying the solution of (4) and assuming a constant thermal profile along the length of an interconnect, the metal temperature, T_m in (7) is given by,

$$T_m = T_{ref} + \Delta T_{self-heating} \tag{8}$$

and,

$$\Delta T_{self-heating} = \left(T_m - T_{ref}\right) = I_{rms}^2\, R\, R_\theta \tag{9}$$

where T_{ref} is the underlying substrate temperature which is typically taken as ~100 ^{0}C, $\Delta T_{self-heating}$ is the temperature rise of the metal interconnect due to the flow of current, R is the interconnect resistance, and R_θ is the effective thermal impedance of the interconnect line to the chip. I_{rms} is the RMS current for a time varying current waveform, or the DC current for a constant current stress. It can be observed from (7) and (9), that as self-heating increases, the metal temperature increases, and hence the EM lifetime decreases exponentially. Therefore, it is important to accurately account for self-heating in (7).

3.2 Self-Consistent Interconnect Design Analysis

In this sub-section the formulation of the self-consistent solutions [14], [15], [18], for allowed interconnect current density is summarized, and then applied to analyze low-k/Cu interconnects in the next sub-section. The $\Delta T_{self-heating}$ in interconnects given by (9) can be written in terms of the RMS current density as,

$$j_{rms}^2 = \frac{\left(T_m - T_{ref}\right) K_{ins}\, W_{eff}}{t_{ins}\, t_m\, W_m\, \rho_m(T_m)} \tag{10}$$

Here t_m and W_m are the thickness and width of interconnect metal line, and $\rho_m(T_m)$ is the metal resistivity at temperature T_m. Note that the thermal impedance R_θ in equation (9) has been expressed as,

$$R_\theta = \frac{t_{ins}}{K_{ins}\, L\, W_{eff}} \tag{11}$$

Here t_{ins} is the total thickness of the underlying dielectric, K_{ins} is the thermal conductivity normal to the plane of the dielectric, and L is the length of the interconnect. In this expression for the thermal impedance, W_{eff} has been modeled as the effective width of the metal line taking quasi-2D heat conduction into consideration from experimental data for high aspect ratio lines [14].

Now, in order to achieve an EM reliability lifetime goal mentioned earlier, we must have the lifetime at any (j_{avg}) current density and metal temperature T_m, equal to or larger than the lifetime value (eg. 10 year) under the design rule current density stress j_0, at the temperature T_{ref}. This value of j_0 is dependent on the specific interconnect metal technology. Therefore we have,

$$\frac{exp\left(\frac{Q}{k_B T_m}\right)}{j_{avg}^2} \geq \frac{exp\left(\frac{Q}{k_B T_{ref}}\right)}{j_0^2} \tag{12}$$

Using the relationship between j_{avg}, j_{peak}, j_{rms}, and r for a rectangular unipolar pulse, ($j_{avg} = r\, j_{peak}$, & $j_{rms} = r^{0.5}\, j_{peak}$) we have after eliminating j_{peak},

$$\frac{j_{avg}^2}{j_{rms}^2} = r \tag{13}$$

Substituting for j_{rms}^2 from (10) and j_{avg}^2 from (12) in (13) we get the self-consistent equation given by,

$$r = j_0^2 \left(\frac{exp\left(\frac{Q}{k_B T_m}\right)}{exp\left(\frac{Q}{k_B T_{ref}}\right)} \right) \frac{t_{ins}\, t_m\, W_m\, \rho_m(T_m)}{\left(T_m - T_{ref}\right) K_{ins}\, W_{eff}} \tag{14}$$

Note that this is a single equation in the single unknown temperature T_m. Once this self-consistent temperature is obtained from (14), the corresponding maximum allowed j_{rms} and j_{peak} can be calculated from (10) and the current density relationships given above. The self-consistent equation given by (14) for unipolar pulses is also valid for more general time varying waveforms with an effective duty cycle r_{eff} [19].

3.3 Impact of New Materials on Current Density Limits

Allowed interconnect current densities are expected to be strongly influenced by low-k materials, which cause increased Joule-heating due to their lower thermal conductivity [12], [20]. Therefore we begin by analyzing the effect of introducing new interconnect and dielectric materials on allowable current density limits. In Figure 2 the self-consistent values of j_{rms} and j_{peak} are plotted as a function of the duty cycle

r, for different dielectrics. It can be observed that j_{rms} and j_{peak} decrease significantly as dielectrics with lower thermal conductivity are introduced. For small values of r, j_{rms} varies very slowly with r. This is due to increased Joule-heating.

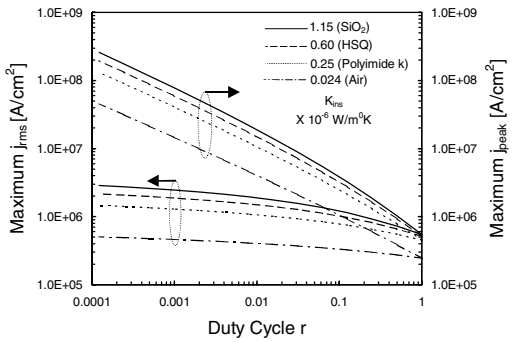

Figure 2: Self-consistent solutions for maximum allowed values of j_{rms} and j_{peak} for Metal 6 in 0.25-µm technology for different values of K_{ins}. Interconnect metal is Cu with $\rho_m(T_m) = 1.67 \times 10^{-6}$ [1 +6.8 x 10^{-3} $^0C^{-1}$ (T_m - T_{ref})] Ω-cm. The activation energy is assumed to be same as that for AlCu.

In Figure 3 we plot j_{peak} and T_m as a function of r for SiO_2 and air as the dielectric for different values of the design current density, j_{avg}. It can be observed that j_{avg} does not change j_{peak} appreciably, and as r reduces, the increase in j_{peak} with j_{avg} becomes negligible. Furthermore, for a dielectric material with low thermal conductivity such as air, j_{peak} is almost independent of j_{avg}. This indicates that introduction of better interconnect materials becomes increasingly ineffective in increasing j_{peak} for dielectrics with poor thermal properties. Figure 4 summarizes the impact of low-k dielectrics on allowed j_{peak} for both Cu and AlCu interconnect systems. It can be observed that the difference between the maximum allowed j_{peak} for AlCu and Cu interconnects reduces as dielectric materials with poor thermal properties are introduced. For the specific case of air as the dielectric, the j_{peak} values are very similar for AlCu and Cu. This is demonstrated for two different values of j_{avg} for Cu, one value identical to that of AlCu and the other three times higher than this value, which accounts for the improved EM performance in Cu.

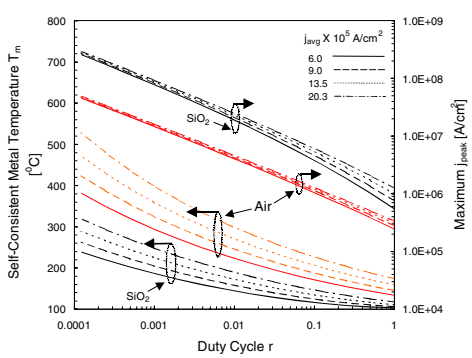

Figure 3: Self-consistent metal temperature and the maximum allowed j_{peak} for Metal 6 in a 0.25-µm Cu technology, as a function of duty cycle for different values of j_{avg} (or j_0) for SiO_2 and air as the dielectric.

4 Interconnect Performance Optimization

4.1 Implications of Performance Optimization on Signal Line Current Densities

As a next step we demonstrate a methodology for computing current density from performance considerations only [14], [15]. Consider an interconnect of length l between two buffers. The schematic representation is shown in Fig. 5(a). Fig. 5(b) shows an equivalent RC circuit for the system. The voltage source (V_{tr}) is assumed to switch instantaneously when voltage at the input capacitor (V_{st}) reaches a fraction x, $0 \le x \le 1$ of the total swing. Hence the overall delay of one segment is given by:

$$\tau = b(x) R_{tr} \left(C_L + C_P\right) + b(x) \left(c R_{tr} + r C_L\right) l + a(x) r c l^2 \qquad (15)$$

where $a(x)$ and $b(x)$ only depend on the switching model, i.e., x. For instance, for x=0.5, a=0.4 and b=0.7 [21].

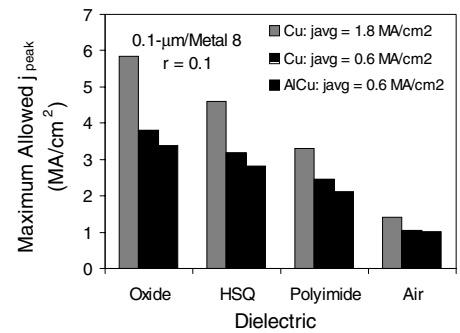

Figure 4: Comparison of the maximum allowed j_{peak} values for AlCu and Cu lines with two different j_{avg} values for Metal 8 of a 0.10-µm technology shown for different dielectric materials.

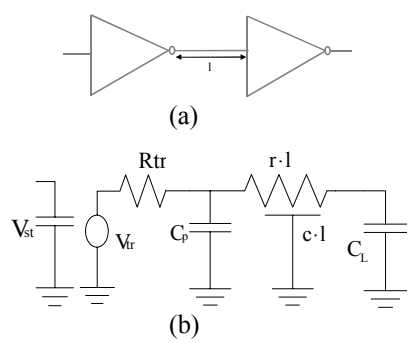

Figure 5: (a) An interconnect of length l between two buffers (b) The equivalent RC circuit. V_{st} is the voltage at the input capacitance that controls the voltage source V_{tr}. R_{tr} is the driver transistor resistance, C_p is the output parasitic capacitance and C_L is the load capacitance of the next stage, r and c are the interconnect resistance and capacitance per unit length respectively.

If r_0, c_0 and c_p are the resistance, input and parasitic output capacitances of a minimum sized inverter respectively then R_{tr} can be written as r_0 / s where s is size of the inverter in multiples of minimum sized inverters. Similarly $C_P = s\, c_p$ and $C_L = s\, c_0$. If the total interconnect of length L is divided into n segments of length $l = L/n$, then the overall delay is given by,

$$T_{delay} = n\tau = \frac{L}{l} b(x) r_0 \left(c_0 + c_p\right) + b(x)\left(c\frac{r_0}{s} + src_0\right)L + a(x) rclL \qquad (16)$$

It should be noted in the above equation that s and l appear separately and therefore T_{delay} can be optimized separately for s and l. The optimum values of l and s are given as:

$$l_{opt} = \sqrt{\frac{b(x) r_0 \left(c_0 + c_p\right)}{a(x) rc}} \qquad (17)$$

$$s_{opt} = \sqrt{\frac{r_0\, c}{r\, c_0}} \qquad (18)$$

Note that s_{opt} is independent of the switching model, i.e., x.

Since, for deep sub-micron technologies, a significant fraction of interconnect capacitance, c, is contributed by coupling and fringing capacitances to neighboring lines, a full 3D-capacitance extraction using SPACE3D [22] for signal lines at various metal levels was used to obtain the values of c for SPICE simulations.

This inverter-interconnect structure is used as a delay stage in a multi-stage ring oscillator and the current waveforms and current densities along the interconnect are be obtained. In practice, the input capacitance C_L of the inverter is almost constant but the output resistance R_{tr} and output parasitic capacitance C_P are bias dependent and therefore change during the output transition. Therefore accurate values of optimal interconnect length and buffer size need to be determined by SPICE simulation. For this, we take advantage of the fact that the optimal interconnect length does not depend on the buffer size. Therefore, we first set the buffer size to an appropriate value and sweep the interconnect length and find the optimum length which minimizes the ratio of the ring oscillator stage delay and interconnect length. Using this optimum length, we subsequently sweep buffer sizes and find the optimum buffer size, which minimizes the stage delay. This allows us to obtain the values of l_{opt} and s_{opt} taking into account the bias dependence of transistor resistances and capacitances and the switching model.

This analysis is carried out for every metal layer in the two technologies under study. Note that due to the distributed nature of the interconnect, the maximum current density occurs close to the buffer output. Hence, we need to verify whether this maximum current density, which is obtained from performance considerations ($j_{-performance}$) only, also meets the EM current density limits ($j_{-reliability}$) obtained earlier using the self-consistent approach.

The interconnect current waveform in the metal lines for 0.25-µm and 0.1-µm technologies, as obtained from SPICE simulations, were found to be bipolar as expected. It should be noted that (7) represents the EM lifetime reliability equation for a unipolar pulse or dc current. The EM lifetime under bipolar stress conditions is known to be higher than that under the unipolar case. In this work, the unipolar EM lifetime reliability equation is used as a worst case limit.

Also, the relative rise and fall skew was found to be same across both technologies. From our simulations it was observed that drivers and interconnects optimized using (17) and (18), maintain good slew rates for rising and falling transitions for all the metal layers and across both technologies with an effective duty cycle ($r_{eff} = j_{avg}^2 / j_{rms}^2$) of 0.12 ± 0.01.

Since for the signal lines the current waveform is symmetric and bipolar, j_{avg} and j_{rms} are computed over half the time-period to obtain r_{eff}. In the above analysis it was assumed that the line capacitance per unit length is constant. Furthermore, it is observed from simulation and can be shown that the interconnect current remains almost the same if the load capacitance of a buffer changes. The rise and fall time will get affected significantly but the interconnect current does not change appreciably if the buffer drive strength is unchanged. However, for the slower transition, r increases and for the faster transition, r decreases (typical values observed in simulation are 0.2 for slower transition and 0.05 for the faster transition).

4.2 Comparisons between Reliability and Performance Based Current Density Limits

Figure 6 shows the comparison of $j_{avg\text{-}performance}$ with the values of $j_{avg\text{-}reliability}$ for isolated and realistic multilevel structures in a 0.1-µm technology. For the isolated line it can be observed that $j_{avg\text{-}performance}$ is always lower than $j_{avg\text{-}reliability}$ for all the dielectrics. However, for low-k dielectric materials we find that the difference between $j_{avg\text{-}reliability}$ and $j_{avg\text{-}performance}$ reduces. This is due to the fact that these dielectrics have lower thermal conductivity than oxide which leads to greater interconnect Joule-heating and therefore lower allowable current densities.

For the more realistic structures (that involve thermal coupling), it can be observed that the $j_{avg\text{-}performance}$ values still remain lower than the $j_{avg\text{-}reliability}$ values even after thermal coupling is taken into account. However, in this simulation it is assumed that r_{eff} = 0.12 at all times, which can actually increase (to ~ 0.2), if neighboring lines switch in the opposite direction as discussed earlier. This increase in r_{eff} will further lower $j_{avg\text{-}reliability}$. Hence, impact of switching states of signal lines on allowed current density design rules is important.

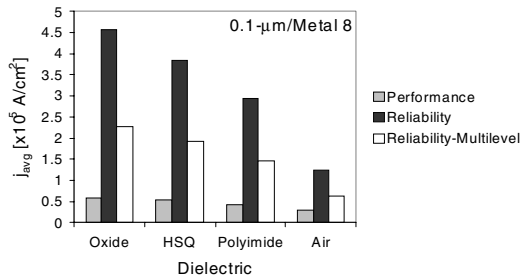

Figure 6: Comparison of j_{avg} values obtained from reliability and performance considerations for Metal 8 of a 0.1-µm Cu technology. r_{eff} = 0.12. The effect of thermal coupling in realistic multilevel interconnect arrays on the j_{avg} values is also shown here for various dielectric materials.

5 Analysis of Non-Uniform Chip temperature on Interconnect Performance

5.1 Non-uniform Temperature Dependent Delay

Consider an interconnect with length L and uniform width w that is driven by a driver of output resistance R_d and terminated at a load with capacitance C_L, with partitioned n equal segments, each with length Δx. Using Elmore delay [23], the delay D of a signal passing through the line can be written as follows:

$$D = R_d((\sum_{i=1}^{n} c_0(x_i)\Delta x) + C_L) + \sum_{i=1}^{n} r_0(x_i)\Delta x(\sum_{j=i}^{n} c_0(x_j)\Delta x + C_L) \qquad (19)$$

where $c_0(x)$ and $r_0(x)$ are the unit length capacitance and unit length resistance at location x, respectively. As the number of the partitions approaches infinity we can rewrite the Elmore delay as follows:

$$D = R_d(C_L + \int_0^L c_0(x)dx) + \int_0^L r_0(x)(\int_x^L c_0(\tau)d\tau + C_L)dx \qquad (20)$$

The third integral in (20) represents the downstream capacitance seen by the interconnect from location x. In an interconnect experiencing temperature profile $T(x)$ along its length, resistance will change linearly with temperature as

$$r_0(x) = \rho_0(1 + \beta \cdot T(x)) \qquad (21)$$

where ρ_0 is the unit length resistance at 0 °C, and β is the temperature coefficient of resistance (1/°C), assuming that unit length capacitance does not change with temperature variations along the interconnect length (which is usually a true assumption). We also assume that the temperature distribution inside the driver is uniform under steady-state condition. Hence the R_d is going to be constant at the chosen operating temperature of the cell. We can simplify (20) to the following [24]:

$$D = D_0 + (c_0 L + C_L)\rho_0\beta\int_0^L T(x)dx - c_0\rho_0\beta\int_0^L xT(x)dx \qquad (22)$$

where:

$$D_0 = R_d(C_L + c_0 L) + (c_0\rho_0\frac{L^2}{2} + \rho_0 L C_L) \qquad (23)$$

D_0 is the Elmore delay (at 0 °C) when the effect of temperature on the line resistance is neglected. Consider circuit parameters for *AlCu* interconnects with β=3E-03 (1/°C) and using r_{sh}=0.077(Ω/sq) at the reference room temperature (25 °C) and c_{sh}=0.268(fF/µm) as the unit sheet resistance and the unit length capacitance, respectively. In an interconnect with w=0.32 µm, R_d=10 Ω and C_L=1000 fF, for each 20

degree increase in the line temperature, there is roughly a 5 to 6 percent increase in the Elmore delay for a long global line (L>2000 µm). In this calculation we used a uniform thermal profile along the interconnect (the worst case scenario for delay degradation). Figure 7 shows the interconnect delay degradations for different wire lengths.

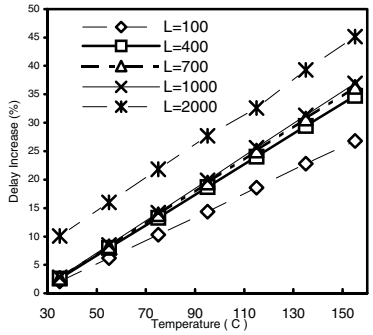

Figure 7: Delay degradation due to the increase of the temperature as a result of a uniform substrate thermal profile and Joule heating.

In reality, and especially for long global lines, the thermal profile along the length of an interconnect is non-uniform, as mentioned earlier. To understand the importance of considering the effect of non-uniform temperature on the delay, assume an interconnect whose two ends are at different temperatures and whose profile between the two ends is modeled by an exponential distribution $T(x)=a.exp(-bx)$ with parameters a and b (Figure 8). Observing the behavior of the line under exponential thermal profiles is important in the sense that most of the solutions of the heat diffusion equation (4) usually have an exponential component.

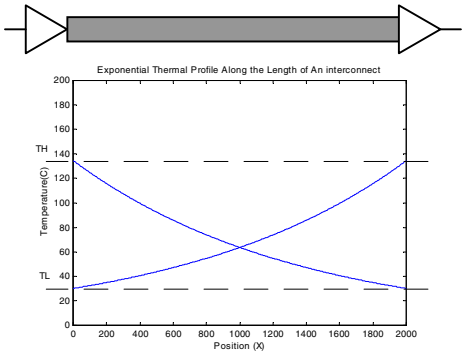

Figure 8: Schematic of exponential thermal profiles along an interconnect.

We apply two different thermal profiles $T_1(x)$ and $T_2(x)$ along the length of the interconnect. Figure 9 compares the delay degradation in the presence of $T_1(x)$ and $T_2(x)$ in two different wire lengths, 1000 µm and 2000 µm, with identical electro-thermal characteristics as mentioned above. In both cases the lower bound temperature is kept constant at 30 °C. By increasing the upper bound value (x-axis of Figure 9) for these functions, it can be observed that using $T_2(x)$ causes less delay increase under the same conditions than when using $T_1(x)$. This shows that the assumption of a constant temperature along the wire (with peak-value) can introduce a large error in planning wire routings and clock-skew analysis. The above observation also demonstrates that if we have the choice, choosing the thermal profile $T_2(x)$ over $T_1(x)$ is preferable.

An explanation for the above behavior is that, from the resistance point of view, fluctuations of the temperature along the line are equivalent to wire sizing with uniform resistance. In sections with higher temperature, the wire can be modeled as a narrower wire, and in sections with lower temperature the wire acts like a wider uniform resistance wire. As a result, an increasing thermal profile is equivalent to a decreasing sizing profile for a uniform resistance wire, which is known to give better delay than that with an increasing sizing profile [25].

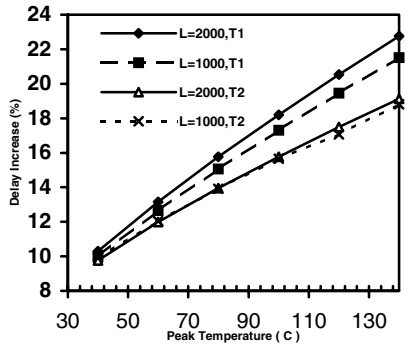

Figure 9: Degradation in interconnect performance caused by $T_1(x)$ and $T_2(x)$ (cf. Figure 8).

5.2 Non-uniform Temperature-Dependent Clock Skew

In addition to the performance degradation introduced by increasing temperature in the interconnect (which causes the effective signal delay to worsen), the non-uniform thermal profile along upper-layer interconnects has a major effect on the skew of the clock signal net. The goal of the clock signal distribution network is to maintain a zero (or near-zero) skew through it. To ensure zero-skew clock distribution, a symmetric H-Tree structure or a bottom-up merging technique can be used [26], [27]. For simplicity and without loss of generality, for our analysis we consider the H-Tree clock topology consisting of trunks (vertical stripes) and branches (horizontal stripes) as depicted in Figure 10. In general, the top-level segments of the tree are wider than the lower-level segments. Furthermore, the top-level global segments of the tree are assigned to the upper metal layers and low-level local segments are routed using the lower metal layers.

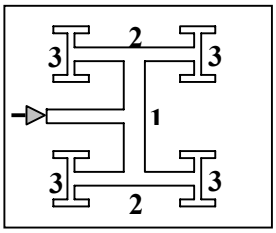

Figure 10: A symmetric H-Tree clock distribution net.

The problem arises from the fact that trunk 1 and branches 2 of the H-Tree are long. Hence, they are exposed to the thermal non-uniformities in the underlying substrate. Such non-uniformity results in different signal delays at the two ends of trunk 1 and branches 2 of the H-Tree; hence there will be a non-zero skew along the tree. The temperature effects, therefore, result in a scenario where the H-tree symmetry cannot guarantee the zero skew. If, for example, trunk 1 experiences a non-uniform thermal profile, the clock driver must be connected to this segment at a place other than the center of the segment. This also suggests that during a bottom-up binary merge construction of the clock tree [26], the actual temperature-dependent delay must be considered. Having more than a 30 °C thermal gradient in some designs, justifies the importance of this kind of analysis. Notice that we consider the *steady-state* thermal profile of the substrate. Even though the dynamic behavior of the chip causes transient changes in the cell switching activities, because of the large time constant for the temperature propagation in the substrate (around a few *ms* [16]), the locations of the hot spots are, in fact, quite stable.

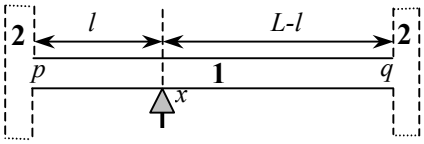

Figure 11: Schematic of a minimum-skew clock signal insertion for an interconnect with a non-uniform temperature profile.

Table 1: Comparison between different thermal profiles and their effects on clock skew.

Thermal Profile	Parameters	*l**	*Normalized Skew %*
$T(x) = ax + b$, $a = \frac{T_H - T_L}{L}$, $b = T_L$	T_H=170, T_L=90	1042	5.42
	T_H=170, T_L=110	1032	3.98
	T_H=170, T_L=130	1021	2.65
	T_H=170, T_L=150	1012	1.29
$T(x) = a \cdot e^{-bx}$, $a = T_H$, $b = \frac{1}{L}\ln(\frac{T_H}{T_L})$	T_H=170, T_L=90	957.5	5.24
	T_H=170, T_L=110	968.66	3.63
	T_H=170, T_L=130	979.5	2.40
	T_H=170, T_L=150	989.7	1.19
$T(x)=T_{max} \cdot e^{-\frac{-(x-\mu)^2}{2\sigma^2}}$	μ=2000, σ=1000	1210	7.78
	μ=1000, σ=400	1000	0.0
	μ=500, σ=400	827	10.7
	μ=300, σ=700	911	9.57

Consider the global trunk 1 in the H-Tree depicted in Figure 11. The goal is to find the division point *x* along the length of the segment (*L*) such that when the clock signal driver is connected to that point, the delay at the two ends of the trunk 1 is the same. This will in turn ensure the minimal effect of the non-uniform temperature gradient on the skew. Assume an interconnect thermal profile *T(x)* along the length *L* of trunk 1. By using the delay model described in III, we can write the propagation delay from the source to the two ends of the trunk. By doing so and assuming balanced loads at the two ends *p* and *q* of the trunk and using (22), the optimum length l^* for ensuring zero clock skew can be obtained by solving the following equation:

$$\beta \int_0^{l^*} T(x)\,dx + l^* - A = 0 \tag{24}$$

where *A* is a constant and can be written as follows:

$$A = \frac{1}{Lc_0 + C_L}\left(\frac{L^2 c_0}{2} + LC_L + \beta(Lc_0 + C_L)\int_0^L T(x)dx - c_0\beta\int_0^L xT(x)\,dx\right) \tag{25}$$

Given circuit parameters *L*, C_L, c_0, β, and *T(x)*, we can easily compute the constant *A* and solve (24) to obtain the optimum position for the clock signal connection to the net segment. From (24) and (25), it is seen that with a constant thermal profile *T(x)* along the length of interconnect, we can guarantee a zero skew by connecting the clock signal at *l=L/2*. In fact, even a non-uniform, but *symmetrical* thermal profile with the symmetry axis at *l=L/2* will result in a zero clock skew when the driver is connected to the middle of the line. From (24), we can also see that a gradually decreasing (increasing) thermal profile along the length of the line from *0* to *L* (from *p* to *q*), results in the optimum length l^* to be less than (greater than) *L/2*.

We now examine the behavior of temperature-dependent clock skew for a 2000 μm line with identical electro-thermal characteristics as those in Section 5.1, by applying three different interconnect thermal profiles. More precisely, we will consider the effects of linear, exponential, and normal (Gaussian distribution with constant peak amplitude) thermal profiles on the clock skew. Since the global clock lines are thermally long, we neglect the thermal effects of vias/contacts at the junction of the interconnect with the driver/receiver. In the first two cases, different scenarios based on high temperature levels (T_H °C) and low temperature levels (T_L °C) have been examined (Table 1). Column 3 shows the value of l^* at which, by inserting the signal to the H-Tree segment, a zero clock skew is guaranteed. The reported normalized skew percentage in column 4 represents the ratio of the clock skew when *l=L/2* over the delay from the driver to any endpoint of the interconnect when $l=l^*$. The third set of thermal profiles uses a constant-peak amplitude normal distribution with peak T_{max} (°C) at 100 °C, mean μ(μm), and standard deviation σ (μm), which approximates the behavior of a hot spot on the substrate. As this profile is symmetric, by applying a distribution with median *L/2*, the zero skew is guaranteed. Moving the hot spot along the length of the line clearly increases the skew. It is clear from Table 1 that neglecting the effects of thermal profiles on the delay fluctuations increases the skew by as much as 10 percent. The above discussion suggests that for a given thermal profile *T(x)*, one can adjust the length of *l* using (24) and (25) to maintain a zero clock skew. The circuit designer can place the cells such that the hot spots have a symmetrical position relative to the higher-level segments of the clock tree or can route the clock tree such that the higher level segments are symmetrical relative to the underlying hot spots. Because the number of these high-level clock segments is small, it is feasible to adjust the position of the clock tree segment or the cell placement over the substrate to maintain a nearly symmetric thermal profile along the clock segments [28], [29].

6 Thermal Effects under High-Current Stress Conditions

Apart from normal circuit conditions, ICs also experience high-current stress conditions, the most important of them being electrostatic discharge (ESD), which causes accelerated thermal failures [30]. Semiconductor industry surveys indicate that ESD is the largest single cause of failures in ICs. ESD is a high-current short-time scale phenomena that can lead to catastrophic open circuit failures and to latent damage [31]. Interconnect failure due to ESD is becoming an important issue as VLSI scaling continues and number of I/O pins increases. As a consequence of this increase in I/O pins, package floor planning is changing from peripheral package connections to array grids in order to accommodate the increased I/O pin count [32]. In the array architecture, the interconnect widths between the external pads and the ESD structures must decrease to preserve chip wirability and to prevent timing delays in critical paths and in the receiver and driver networks. This trend can increase the susceptibility of interconnects to ESD failure. Furthermore, technology scaling and the transition to new interconnect and dielectric materials necessitates a growing need to comprehend the high-current behavior of these structures and analyze their failure mechanisms in order to provide robust design guidelines.

In [7] it has been shown that the critical current density for causing open circuit metal failure in AlCu interconnects is ~ 60 MA/cm^2. A short-time scale high-current failure model for designing robust interconnects to avoid thermal failure under high peak current conditions is also formulated in [7]. This model can be used to determine the critical current for open circuit metal failure in terms of the pulse width and the line width. A study from IBM has shown that the model can also be applied to design damascene Cu interconnects [32]. As mentioned earlier, interconnects can also suffer latent damage if the lines resolidify after melting; this has been shown to degrade the EM lifetime [8]. The model presented in [7] can also be used to avoid this latent reliability hazard. These interconnect design rules must be obeyed for high-current robustness.

7 Summary

This paper provided an overview of various thermal issues in high-performance VLSI with especial attention to their implications for interconnect reliability and performance. It showed that VLSI scaling is causing increased thermal effects in high-performance ICs. It was demonstrated that coupled analysis of EM reliability, thermal effects, and

interconnect performance optimization is necessary to quantify their impact on current density limits and to understand various tradeoffs between technology, reliability, and performance issues. A methodology that allows the determination of both reliability and performance based current density limits was discussed. This technique can be effectively applied to study the impact of technology scaling and performance optimizations on interconnect reliability. For low-k/Cu interconnect systems, it was shown that as long as point-to-point interconnect performance can be optimized, EM design limits for those signal lines will be satisfied.

Additionally, it was demonstrated how chip level thermal effects, such as non-uniform substrate temperature, can have a significant impact on large-scale circuit optimization techniques, including the clock-skew minimization scheme, and can influence other physical design problem formulations. Finally, high-current interconnect design rules for ESD and I/O circuits were also examined.

In conclusion, thermal issues are a growing problem for high-performance VLSI and detailed electrothermal modeling of both device and interconnects would be necessary to understand the tradeoffs between performance, reliability and cost. A Recent work has shown that in the near future advancement in chip packaging and cooling technology will be necessary for high-performance ICs [3]. Recently dummy thermal vias have been shown to be effective in alleviating thermal problems in multilevel interconnects [33]. Similar approach can be employed to alleviate thermal problems at the chip level.

8 Acknowledgments

The authors would like to acknowledge support from the MARCO Interconnect Focus Center at Stanford and the SRC.

9 References

[1] A. P. Chandrakasan and R. W. Brodersen, *Low Power Digital CMOS Design*, Kluwer Academic Publishers, 1995.

[2] *International Technology Roadmap for Semiconductors*- ITRS, 1999.

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

[4] K. Banerjee, "Thermal effects in deep sub-micron VLSI interconnects," *Tutorial Notes*, *IEEE International Symposium on Quality Electronic Design*, 2000.

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

[6] B. K. Liew, N. W. Cheung, and C. Hu, "Projecting interconnect electromigration lifetime for arbitrary current waveforms," *IEEE Trans. Electron Devices*, vol. 37, pp. 1343-50, 1990.

[7] K. Banerjee, A. Amerasekera, N. Cheung, and C. Hu, "High-current failure model for VLSI interconnects under short-pulse stress conditions," *IEEE Electron Device Lett.,* vol. 18, No. 9, pp. 405-407, 1997.

[8] K. Banerjee, D. Y. Kim, A. Amerasekera, C. Hu, S. S. Wong, and K. E. Goodson, "Microanalysis of VLSI interconnect failure modes under short-pulse stress conditions," *IRPS*, 2000, pp. 283-288.

[9] J. Ida et al., "Reduction of wiring capacitance with new low dielectric SiOF interlayer film for high speed/low power sub-half micron CMOS," *Tech. Dig. VLSI Symp.*, 1994, pp. 59-60.

[10] *MRS Bulletin*, October 1997.

[11] B. Shieh, K. C. Saraswat, J.P. McVittie, S. List, S. Nag, M. Islamraja, and R.H. Havemann, "Air-Gap formation during ILD deposition to lower interconnect capacitance," *IEEE Electron Device Lett*., vol. 19, no. 1, pp. 16-18, 1998.

[12] K. Banerjee, A. Amerasekera, G. Dixit, and C. Hu, "The effect of interconnect scaling and low-k dielectric on the thermal characteristics of the IC metal," *Tech. Dig. IEDM*, 1996, pp. 65-68.

[13] J. Culetu, C. Amir, and J. McDonald, "A practical repeater insertion method in high speed VLSI circuits," *ACM Design Automation Conference*, 1998, pp. 392-395.

[14] K. Banerjee, A. Mehrotra, A. Sangiovanni-Vincentelli, and C. Hu, "On thermal effects in deep sub-micron VLSI interconnects," *ACM Design Automation Conference,* 1999, pp. 885-891.

[15] K. Banerjee, A. Mehrotra, W. Hunter, K. C. Saraswat, K. E. Goodson, and S. S. Wong, "Quantitative projections of reliability and performance for low-k/Cu interconnect systems," *IRPS*, 2000, pp. 354-358.

[16] Y. Cheng, C. Tsai, C. Teng, S. Kang, *Electrothermal Analysis of VLSI Systems*, Kluwer Academic Publishers, 1st ed., 2000.

[17] H.A. Schafft, "Thermal Analysis of Electromigration Test Structures," IEEE Trans. on Electron Device, vol.Ed-34, No.3, pp.664-672, 1987.

[18] W. R. Hunter, "Self-consistent solutions for allowed interconnect current density – Part I: Implications for technology evolution," *IEEE Trans. Electron Devices*, vol. ED-44, pp. 304-309, 1997.

[19] W. R. Hunter, "Self-consistent solutions for allowed interconnect current density – Part II: Application to design guidelines," *IEEE Trans. Electron Devices*, vol. ED-44, pp. 310-316, 1997.

[20] K. E. Goodson and Y. S. Ju, "Heat conduction in novel electronic films," *Annu. Rev. Mater. Sci*., 29: pp. 261-293, 1999.

[21] R. H. J. M. Otten and R. K. Brayton, "Planning for performance," *Proc. 35th Annual Design Automation Conference*, 1998, pp. 122-127.

[22] "Physical design modeling and verification project (SPACE)," *http://cas.et.tudelft.nl/research/space.html*

[23] W.C. Elmore, "The Transient Response of Damped Linear Network with Particular Regard to Wide-Band Amplifier," *Journal of Applied Physics*, vol.19, pp.52-63, 1948.

[24] A. H. Ajami, K. Banerjee, M. Pedram, and L. P.P.P. van Ginneken, "Analysis of Non-Uniform Temperature-Dependent Interconnect Performance in High-Performance ICs," to appear in *38th ACM Design Automation Conference*, 2001.

[25] C-P. Chen, *et al.*, "Optimal wire-sizing formula under the Elmore delay model," *Proc. Design Automated Conference*, 1996, pp. 487-490.

[26] T.H. Chao, Y.C. Hsu, J.M. Ho, K.D. Boese, A.B. Kahng, "Zero skew clock routing with minimum wirelength," *IEEE Transaction on Circuits and Systems-II*, vol. 39, No. 11, pp. 799-814, 1992.

[27] P. Zarkesh-Ha, T. Mule, J.D. Meindl, "Characterization and modeling of clock skew with process variation," *Proc. Custom Integrated Circuits Conf.*, 1999, pp. 441-444.

[28] A.H. Ajami, M. Pedram, K. Banerjee, "Effects of non-uniform substrate temperature on the clock signal integrity in high performance designs," *to appear in Proc. IEEE Custom Integrated Circuits Conf.*, 2001.

[29] A.H. Ajami, K. Banerjee, and M. Pedram, "Non-uniform chip-temperature dependent signal integrity," *to appear in Symposium on VLSI Technology.*, 2001.

[30] C. Duvvury and A. Amerasekera, "ESD: A pervasive reliability concern for IC technologies," *Proc. of the IEEE,* Vol. 81, No. 5, pp. 690-702, 1993.

[31] C. Duvvury and A. Amerasekera, "State-of-the-art issues for technology and circuit design of ESD protection in CMOS ICs," *Semiconductor Science. and Tech*., pp. 833-850, 1996.

[32] S. H. Voldman, "ESD robustness and scaling implications of aluminum and copper interconnects in advanced semiconductor technology," *Proc. EOS/ESD Symposium*, 1997, pp. 316-329.

[33] T-Y Chiang, K. Banerjee, K. C. Saraswat, "Effect of via separation and low-k dielectric materials on the thermal characteristics of Cu interconnects," *Tech. Dig. IEDM*, 2000, pp. 261-264.