Modeling and Analysis of Non-Uniform Substrate Temperature Effects in High Performance VLSI

Amir H. Ajami[1], Kaustav Banerjee[2], and Massoud Pedram[1]

[1]Dept. of EE-Systems, University of Southern California, Los Angeles, CA 90089

(aajami@usc.edu, massoud@ceng.usc.edu)

[2]Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 (kaustav@ece.ucsb.edu)

Abstract

A non-uniform temperature profile for the chip substrate in high-performance ICs can significantly impact the performance of the global on-chip interconnects. This paper presents a detailed modeling and analysis of the interconnect performance degradation due to the non-uniform temperature profiles that are encountered along the metal connections as a result of the thermal gradients in the underlying Silicon substrate. More precisely, a non-uniform temperature-dependent distributed RC interconnect delay model is proposed. The model is applied to a wide variety of interconnect layouts and substrate temperature distributions to quantify the impact of such thermal non-uniformities on signal integrity issues including speed degradation and clock skew fluctuations. Subsequently, a new thermally dependent zero-skew clock routing methodology is presented.

Keywords: clock skew, Elmore delay, global interconnects, signal integrity, substrate-temperature gradient.

1 Introduction

CMOS device technology has scaled rapidly for nearly three decades and has come to dominate the electronics world. Because of this scaling, CMOS circuits have become extremely dense and operate at ever-increasing clock frequencies. Consequently, power dissipation and substrate temperature have become major design considerations. Although industry projections call for at least another 10 years of progress, making progress will be difficult, and will likely to be significantly constrained by power dissipation and heat generation. Thermal issues are rapidly becoming one of the most challenging problems in high-performance IC design due to aggressive device scaling trends [1],[2]. Hence, thermal management has become an essential step in the design and development of future generations of high-performance microprocessors, integrated network processors, and systems-on-a-chip. At the circuit level, thermal effects have important implications for both circuit performance and reliability [3],[4],[7].

As the minimum feature size of the CMOS fabrication process continues to scale down, the performance of the CMOS integrated circuits has become dominated by the global interconnect lines [5],[6]. Aggressive scaling in CMOS switching speeds causes an increase in the average current density. This increase has had a big impact on interconnect reliability and performance. The interconnect reliability degradation is mainly due to the electromigration (EM) phenomenon, although Joule heating (self-heating) is another important contributor to the interconnect reliability degradation. More precisely, the self-heating effect results in a temperature rise in the interconnect lines, which in turn causes an exponential reduction of the (EM-induced) interconnect time-to-failure [4]. The interconnect performance degradation is primarily due to the fact that the resistivity of a metal line increases linearly with its temperature.

Although extensive work has been done to determine the chip temperature and predict the effect of temperature on EM reliability of interconnects [3],[4],[8],[10], few efforts have focused on analyzing the effect of temperature on performance of interconnects [3]. Recent work indicates that in high performance ICs the peak chip temperature can rise up to 140 ºC in 0.13 μm technology feature size and is expected to rise to a much higher level for future technologies [11]. Such a temperature rise can significantly increase the interconnect resistance, which would in turn increase the signal propagation delay in the interconnect line. In some cases, this effect has been shown to cause timing violations [12]. However, most of the previous works on timing analysis have assumed a *uniform* temperature profile in the substrate. This assumption is, however, invalid. It is well known that large temperature gradients can occur in the substrate of high-performance microprocessor chips. These gradients, for example, may be created due to

“spotty” gate-level switching activity and/or because of the functional blocks are put in different operational modes, e.g., active, standby, or sleep modes [13]. It has been reported in [14] that thermal gradients as high as 40 °C can exist in high-performance microprocessor designs. Dynamic power management (DPM) [15] and clock gating can be major contributors to a non-uniform substrate temperature.

Based on the cell-level power consumption map of the substrate, researchers have provided efficient techniques to obtain a temperature profile of the substrate surface [16]. The existence of such thermal gradients on the substrate results in non-uniform temperature profiles along the lengths of the global interconnect lines running above the substrate, which in turn leads to non-uniform resistance profiles for these interconnect lines. Such non-uniformity in the resistance of global interconnects strongly impacts many aspects of interconnect performance modeling and optimization. In addition, as feature sizes scales to sub-0.1μm dimensions, in spite of an increase in the number of metal layers that will be available in high-performance ICs, the top metal layers will be getting closer to the substrate [8], which will further increase the translation (“coupling”) of thermal gradients from the substrate to the interconnect lines. Clearly, the dependence of interconnect performance on non-uniform temperature distributions along the length of global wires will have a big impact on the solutions to many physical design and layout optimization problems, including clock skew control, wire sizing, layer assignment, crosstalk effects, and buffer insertion. This observation suggests that non-uniform temperature profiles along the interconnect lines should be considered during the design optimization flow and proper steps must be taken to ensure optimal performance [17].

This paper presents a study of non-uniform interconnect thermal profile and its impact on signal integrity in general and clock skew in particular. The paper is organized as follows: Section 2 describes the background for interconnect temperature profile calculation and provides a systematic way of estimating the temperature along the length of an interconnect line in the presence of substrate thermal gradients. The actual boundary conditions and interconnect-via/contact arrangements inside the chip are used to obtain the thermal profiles along the interconnect lines. Section 3 introduces the effects of non-uniform interconnect thermal profile on the Elmore delay. A distributed *RC* interconnect delay model as a function of the line temperature is proposed, and a design methodology for improving the performance in the presence of non-uniform interconnect thermal profiles is presented. Section 4 illustrates the effects of non-uniform interconnect temperature profiles on the clock skew. New design rules are proposed to ensure optimal layout of the temperature-dependent zero-skew clock routing tree. Finally, concluding remarks are made in Section 5. Preliminary versions of this work have been published in [18]and [19].

2 An Analytical Model for the Interconnect Thermal Profile

2.1 General Theory

The temperature distribution as a function of position (*r)* and time (*t*) in a closed structure is governed by the following heat diffusion equation and proper boundary conditions:

$$-\nabla\cdot(-k\nabla T(t,r))+Q(r)=\frac{d}{dt}c_pT(t,r) \qquad (2.1)$$

subject to some defined initial values. T is the time- and position-dependent temperature at each position coordinate r, k is the solid thermal conductivity of the material as a function of temperature (W/(m°C), c_p is the specific heat (J/(kg°C)) of the material constituting the structure, and Q is the position-dependent heat generation rate. In a general multi-layer structure, k and Q are position-dependent, i.e. are functions of r. In a 3-D space (x,y,z), the heat diffusion equation (2.1) in any material can be written as [18]:

$$\frac{\partial}{\partial x}(k\frac{\partial T}{\partial x})+\frac{\partial}{\partial y}(k\frac{\partial T}{\partial y})+\frac{\partial}{\partial z}(k\frac{\partial T}{\partial z})+Q^*=\delta\, c_p\frac{\partial T}{\partial t} \qquad (2.2)$$

where Q^* is the rate of heat generation per unit volume (W/m^3), δ is the solid density (kg/m^3). In general, a boundary condition for solving the diffusion equation (2.2) can be written as follows:

$$k\frac{\partial T}{\partial n_i}+h_i\cdot T=f_i \qquad (2.3)$$

$\partial/\partial n$ denotes the differentiation along the outward-drawn normal at the boundary surface s_i, h_i is the heat transfer function from surface s_i (W/(m^2°C)), and f_i is an arbitrary function of position in the space.

Although the thermal conductivity k is generally a function of temperature and position, due to its rather small variations in the conductors, k is often assumed to be a constant when analyzing VLSI interconnection lines. In addition, the four sidewalls and the top surface of the chip containing the interconnect lines are presumed to be completely insulated (which is generally a valid assumption). This means that the interconnect lines can exchange heat with the external environment only through the bottom face, i.e., the chip substrate, which is in turn connected to the heat-sink. Based on these simplifying assumptions and working under the steady-state condition, the system of heat equation (2.2) and boundary conditions can be reduced as follows:

$$k(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2})+Q^*_{eff}=0 \qquad (2.4)$$

subject to specified initial conditions. Note that Q^{*}_{eff} is the *effective* volumetric heat generation, which also considers the heat *loss* rate per unit volume that addresses the functionality of the boundary condition (heat loss) from the bottom side of the interconnect line. In order to find an exact solution we need to employ a 3-D finite element thermal simulation [7]. On the other had, for a global interconnect line, the length of the line is much larger then its thickness or width, i.e., the thermal gradients along the thickness and width of the interconnect line can be ignored when studying the long interconnects. Consequently, many researchers used a simplified version of (2.4) and employed the 1-D heat diffusion equation to avoid the huge computation time used by FEM simulators [21] while marinating acceptable results. In that case, equation (2.4) can be reduced as follows:

$$\frac{d^2T}{dx^2} = -\frac{Q^{*}_{eff}}{k_m} \tag{2.5}$$

where k_m is the thermal conductivity of the metal. To derive the effective volumetric heat generation Q^{*}_{eff}, consider an interconnect line passing over the substrate as shown in Figure 1. The interconnect line is connected to the substrate through via's at its two ends. The major source of temperature generation in a chip is the power dissipation due to the dynamic and static activity of the cells lying on the substrate. In addition, the power dissipation in the interconnect line is also a source of the heat generation.

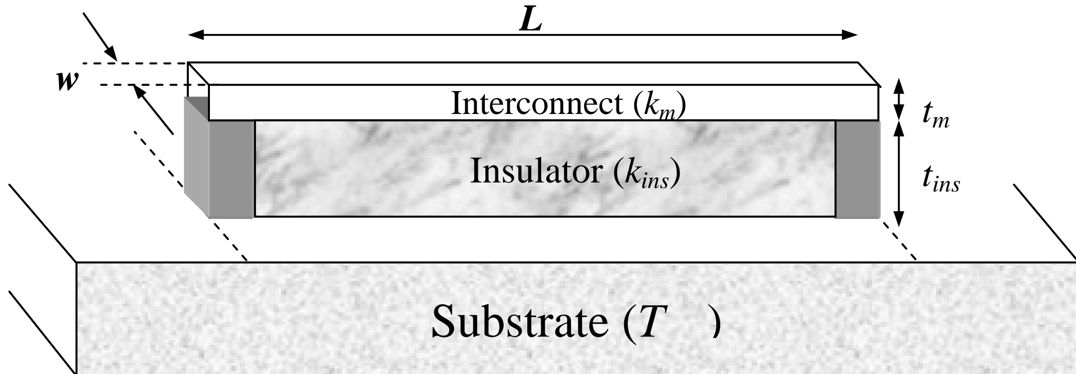

Figure 1. An interconnect line running over the substrate and insulator layer.

For the interconnect line shown in Figure 1 the power dissipation P_g in a partial metal length Δx can be expressed as:

$$P_g(x) = I_{rms}^2 \, \Delta R_E(x) \tag{2.6}$$

where I_{rms} is the root mean square current passing through the line. The electrical resistance of the interconnect line R_E has a linear relationship with its temperature and can be written as follows:

$$R_E(x) = R_0(1 + \beta \cdot T(x)) \quad (2.7)$$

where R_0 is the resistance per unit length at a reference temperature, β is the temperature coefficient of resistance (1/°C), and $T(x)$ is the temperature profile along the length of the interconnect line. Furthermore, initial resistance R_0 can be expressed as:

$$\Delta R_0(x) = \rho_i \frac{\Delta x}{w t_m} \quad (2.8)$$

where ρ_i is the electrical resistivity of the interconnect at the reference temperature, t_m is the interconnect thickness and w is the width of the interconnect. On the other hand, energy loss due to heat transfer between the interconnect and the substrate through the insulator for a partial length Δx is:

$$P_l(x) = \frac{T_{line}(x) - T_{sub}(x)}{\Delta R_T(x)} \quad (2.9)$$

where:

$$\Delta R_T(x) = \frac{t_{ins}}{k^*_{ins} w_m \Delta x} \quad (2.10)$$

$P_l(x)$ is the heat flow from the interconnect to the substrate, T_{line} is the interconnect temperature, T_{sub} is the underlying substrate temperature, R_T is the insulator thermal resistance, and k^*_{ins} is the *effective* insulator thermal conductivity. k^*_{ins} is a shape-dependent parameter which considers the geometry configuration of the heat conducting body on the thermal conductivity. In the case of heat flow by conduction between two identical flat plates with insulated edges, k^*_{ins} is simply the thermal conductivity k_{ins}. For the case of having a rectangular shape parallel to an infinite plate, a simple approximation introduced by Bilotti [22] can be used. This approximation uses a quasi 2-D model where heat flows only through the bottom side and partially by the two sides along length of the rectangular shape (interconnect). In this case, the effective thermal conductivity k^*_{ins} can be expressed as $k_{ins}((1+0.88t_{ins}/w)$ and provides results with 3% accuracy for test cases with $w_m/t_{ins} > 0.4$. However, in deep submicron technologies, the geometrical dimensions of global lines will not satisfy this condition. Hence, using Bilotti's estimation for effective thermal conduction often results in somewhat higher values for the peak temperature in global lines compared to the actual ones. In reality, the heat flows from all sides of the rectangular body (i.e., the interconnect). A more accurate expression for k^*_{ins} was given in [23] where it

takes this factor into account and therefore provides a more accurate expression for the thermal conductivity as follows:

$$k_{ins}^{*} = k_{ins} \cdot \frac{t_{ins}}{w_m} \cdot 1.685 \cdot [\log(1 + \frac{t_{ins}}{w_m})]^{-0.59} \cdot (\frac{t_{ins}}{t_m})^{-0.078} \quad (2.11)$$

Authors in [24] have used this approximation and validated its accuracy with 3-D FEM simulations with negligible error for rectangular-shaped interconnect lines.

Based on the above observations, the net heat energy gain per unit volume is:

$$Q^{*}_{eff} = \frac{P_g - P_l}{w t_m \Delta x} \quad (2.12)$$

Using the simplified heat equation (2.5), the summarized interconnect heat flow equation can be written as follows:

$$\frac{d^2 T_{line}(x)}{dx^2} = \lambda^2 T_{line}(x) - \lambda^2 T_{sub}(x) - \theta \quad (2.13)$$

$$\lambda^2 = \frac{1}{k_m} (\frac{k_{ins}^{*}}{t_m t_{ins}} - \frac{I_{rms}^2 \rho_i \beta}{w^2 t_m^2}) \quad (2.14)$$

$$\theta = \frac{I_{rms}^2 \rho_i}{w^2 t_m^2 k_m} \quad (2.15)$$

where λ and θ are constants in specified technology and interconnect layer assignment. Equation (2.13) and its coefficients will be the basis of our interconnect temperature calculations. Note that in order to have a unique solution for (2.13), we also need to provide two initial conditions. Equation (2.13) shows that the underlying substrate temperature, $T_{ref}(x)$, plays an important role in determining the temperature of the line. This value is usually assumed to be constant throughout the substrate surface. Although this is a valid assumption for the short local interconnects, it is not true in the case of long global lines in the upper metal layers. Because of the different switching activities of various cells on the substrate surface, a non-uniform temperature profile along the substrate surface is inevitable. In this study two cases have been analyzed: 1) uniform thermal profile over the underlying substrate and 2) non-uniform thermal profile over the underlying substrate.

2.2 A Uniform Substrate Temperature Profile

Assume that $T_{ref}(x)$ is constant for all positions along the length of the line. The two initial conditions that are needed to solve (2.13) can be derived using the interconnect line and the via/contact setup. For one

segment of a signal net there are four possible configurations, depicted in Figure 2, based on the location and connection of the via's. Here we examine the routes between substrate and metal 1 and between metal 1 and metal 2. One can easily extend these configurations in the same manner to the other metal layers. We assume that via's get as hot as the layers that are immediately beneath them. In reality (and especially in *AlCu* technology), due to their smaller cross-sectional area and higher electrical resistivity, via's can become much hotter [25] (unless they have been arranged in some sort of via array instead of just one via contact.) In the present analysis, it is assumed that the router uses via arrays wherever it is possible.

Considering Figure 2(a), we see that the two end via's create a thermally conductive path between the metal layer and the substrate. Due to the very small thermal resistivity of via's, we assume that the temperature at the two sides of the metal line is equal to the temperature of the substrate. For example, the initial conditions in Figure 2(a) to solve (2.13) can be written as follows:

$$T(x=0)=T_{sub} \quad , \quad T(x=L)=T_{sub} \qquad (2.16)$$

where $0 \le x \le L$ and T_{sub} is the constant substrate temperature. By solving the homogenous differential equation (2.13) with constant coefficients given by (2.14) and (2.15), the line temperature can be written as follows:

$$T(x)=T_{ref}+\frac{\theta}{\lambda^2}\left(1-\frac{\sinh\lambda x+\sinh\lambda(L-x)}{\sinh\lambda L}\right) \qquad (2.17)$$

Assuming a uniform substrate temperature of 100 °C, the interconnect line thermal profile for global lines corresponding to Figure 2(a) for two different technologies (where parameters provided by ITRS [27]) are depicted in Figure 3. Distance *d* is defined as the heat diffusion length; it is a function of $1/\lambda$ and is strongly dependent on the thickness of the insulator between the metal and the substrate and the effective current density flowing through the metal. Using (2.17) and assuming a constant current density in all metal layers of a signal net, the diffusion length *d* is larger for the higher level metal layers due to their higher underlying insulator thickness. As an example, for an interconnect with an RMS current of 2*mA* in a metal layer with width 0.32 µm and an underlying oxide layer with thickness 1.2 µm, the diffusion length *d* is approximately 40 µm. In addition, the peak value of the temperature in Figure 3 is equal to θ/λ^2. For interconnects whose lengths are comparable to the heat diffusion lengths, the line temperature does not reach the maximum peak value. Using this concept, there are new techniques to make the peak temperature lower by adding extra dummy via's separated by a distance less than the diffusion length.

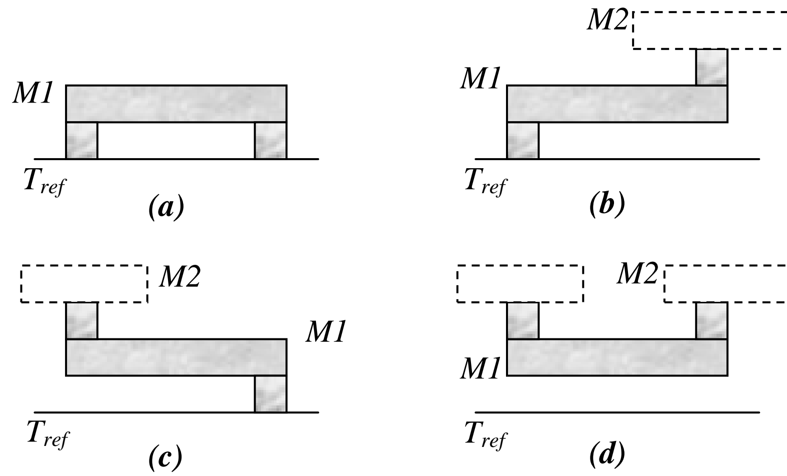

Figure 2. Different configurations of metal lines and via's.

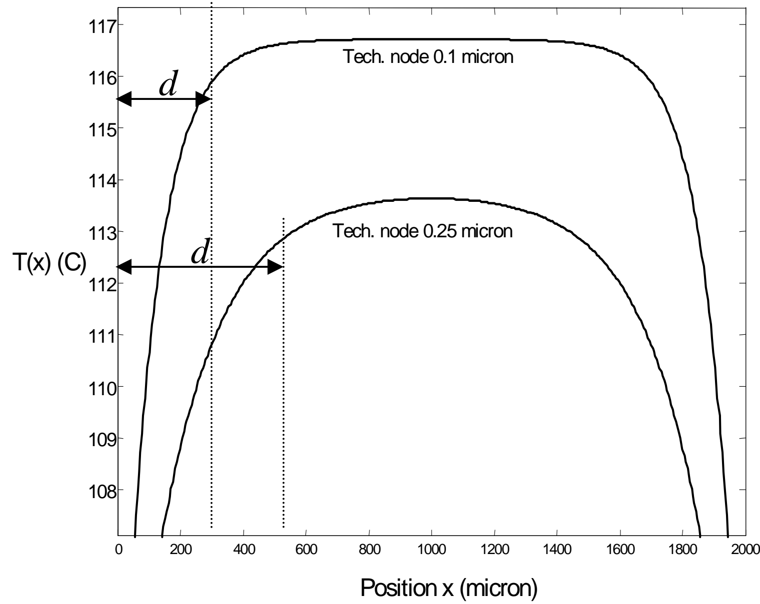

Figure 3. Thermal profile along the length of a 2000 μm long interconnect (Cu) line with a uniform substrate temperature when using interconnect parameters for global metal layers in the 0.1 μm and 0.25 μm technology nodes provided by ITRS [27].

2.3 Non-uniform Substrate Temperature Profile

Considering different switching activities and power consumptions of the cells, substrate thermal profile $T_{sub}(x)$ cannot be a constant value in all locations x on the substrate surface. The quality of extracting the

$T_{sub}(x)$ depends on how accurately one estimates the power consumptions of the cells or macro-cells. Some techniques that are used to find the substrate thermal profile are given in [16]. Due to the duality between thermal and electrical networks, the easiest way to map the substrate thermal profile is to model the substrate as a 3-D grid and solve the system of thermal relations between each two nodes in the grid while considering the packaging and the ambient temperature as additional thermal nodes. A more simplistic technique can be achieved by using a 2-D mesh over the substrate surface (Figure 4). This can be achieved by using the concept of *transfer thermal resistance*. By definition, the transfer thermal resistance $R^T{}_{ij}$ of a location j (in the 2-D mesh or 3-D grid) with respect to a point heat source i can be defined as:

$$R^T{}_{ij} = \frac{T_{ji}}{P_i} \tag{2.18}$$

which is basically the dual of the electrical relationship between current in an electrical resistance and the voltages at its two ends. Using the finite difference method [16], one can easily find the transfer thermal resistance values of all surface nodes with respect to any single source node by sampling the temperature of these nodes due to one unit of dissipated heat at the source node. In general, by using a 3-D grid structure over the substrate we can formulate an $n \times n$ thermal resistance matrix for 3-D nodes by using (2.18), where n is the number of nodes in the grid. Using the transfer thermal resistance matrix $\mathbf{R}^T{}_{n\times n}$, one can easily calculate the temperature distribution $\mathbf{T} = [T_1, T_2, \ldots, T_n]^t$ at each node of the grid due to a given specific heat distribution $\mathbf{P} = [P_1, P_2, \ldots, P_n]^t$, by solving $\mathbf{T} = \mathbf{R}^t \times \mathbf{P}$.

Because this procedure depends on finding the power map of the cells on the substrate, $T_{ref}(x)$ is a design dependent function. For this reason, and for illustration, we use two linear substrate temperature distributions along the length of an interconnect and observe their effects on interconnect temperature $T(x)$ variations. First, we use $T_1(x)=ax+b$ and solve the non-homogeneous differential heat equation (2.14) for the configuration shown in Figure 2(a) with proper initial conditions. The resulting thermal profile along the line can be expressed as:

$$T_1(x) = \frac{\theta}{\lambda^2}(1 - e^{-\lambda x} - \frac{1 - e^{-\lambda L}}{\sinh \lambda L} \sinh \lambda x) + ax + b \tag{2.19}$$

Figure 5 shows the thermal profile in an interconnect using the linear substrate thermal profile $T_1(x)$ (with a gradient from 30 °C to 100 °C).

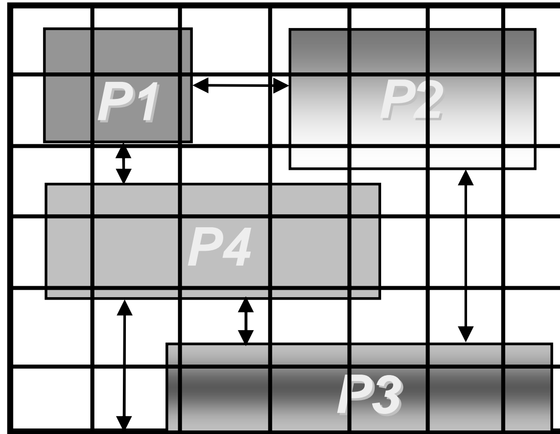

Figure 4. Illustrating the concept of imposing a 2-D mesh on the substrate surface for determining $T_{ref}(x)$, and using the thermal resistance between two adjacent nodes in the mesh by considering the power consumption of each block. One can extend this technique to a 3-D grid model and solve the system of equations to derive the temperature of each node in the 3-D grid.

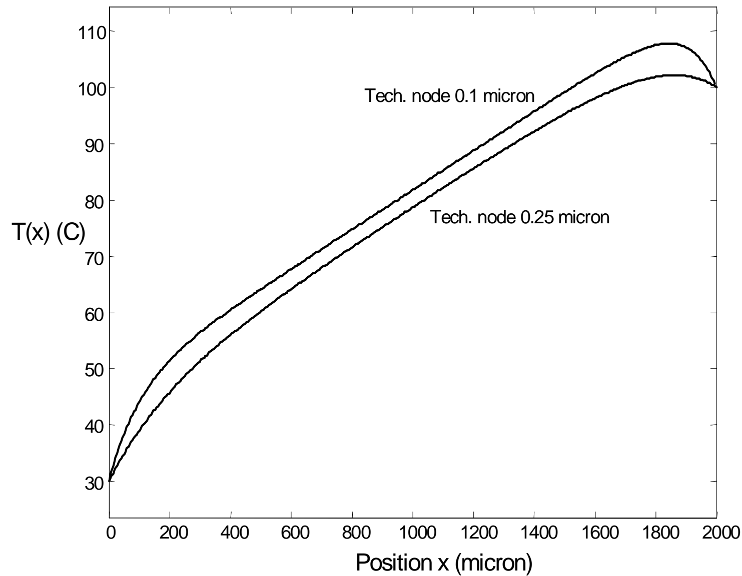

Figure 5. Thermal profile along the length of a 2000 μm long interconnect (Cu) line with a linear substrate thermal profile when using parameters of global wires of 0.1 μm and 0.25 μm technologies provided by the ITRS.

3 A Non-Uniform Temperature-Dependent Interconnect Delay Model

Consider an interconnect with length L and uniform width w that is driven by a driver with on-resistance R_d and junction capacitance C_p terminated by a load with capacitance C_L as depicted in Figure 6.

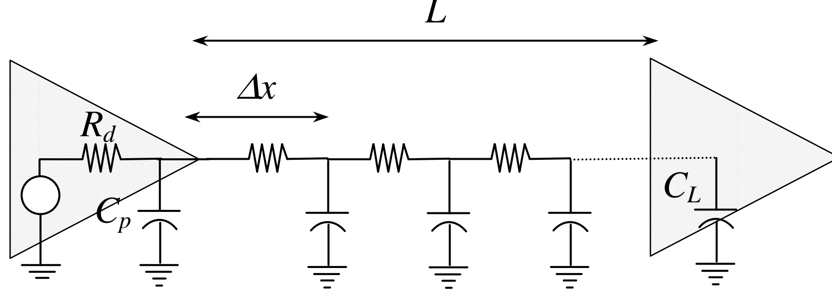

Figure 6. A distributed *RC* interconnect line model driven by source resistance R_d and terminated by load capacitance C_L.

The line is partitioned into n equal segments, each with length Δx. By using the distributed *RC* Elmore delay model, delay D of signal propagation through the line can be written as follows:

$$D = R_d((\sum_{i=1}^{n} c_0(x_i)\cdot\Delta x) + C_L) + \sum_{i=1}^{n} r_0(x_i)\cdot\Delta x\cdot(\sum_{j=i}^{n} c_0(x_j)\cdot\Delta x + C_L) \qquad (2.20)$$

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

$$D = R_d(C_L + \int_0^L c_0(x)dx) + \int_0^L r_0(x)\cdot(\int_x^L c_0(\tau)d\tau + C_L)dx \qquad (2.21)$$

The third integral in (2.21) represents the downstream capacitance seen by the interconnect line from location x. It is assumed that the capacitance per unit length does not change with temperature variations along the interconnect length (which is generally a true assumption). It is also assumed that the temperature distribution inside the driver is uniform under the steady-state condition. Hence the R_d will be constant at the chosen operating temperature of the cell. Using (2.7), We can simplify (2.21) as follows:

$$D = D_0 + (c_0 L + C_L)\rho_0\beta\int_0^L T(x)dx - c_0\rho_0\beta\int_0^L x\cdot T(x)dx \qquad (2.22)$$

where:

$$D_0 = R_d(C_L + c_0 L) + (c_0 \rho_0 \frac{L^2}{2} + \rho_0 L C_L) \qquad (2.23)$$

D_0 is the Elmore delay of the interconnect corresponding to the unit length resistance at reference temperature.

From (2.22) it is clear that in order to calculate the actual temperature-dependent delay we need to compute the area under $T(x)$ and $xT(x)$. To get an idea of how much temperature can affect the degradation of the delay, we assume the worst case scenario by using a uniform thermal profile at some peak temperature over the entire length of the interconnect. Choosing electrical and thermal parameters for *AlCu* interconnects with β=3E-03 (1/°C) and using r_{sh}=0.077(Ω/sq) at room temperature (27°C) and c_{sh}=0.2(fF/sq) as the unit sheet resistance and capacitance, respectively, the variations of Elmore delay with temperature in an interconnect line with w=0.32 μm, R_d=10Ω, and C_L=1000fF for different lengths in μm are summarized in Figure 7.

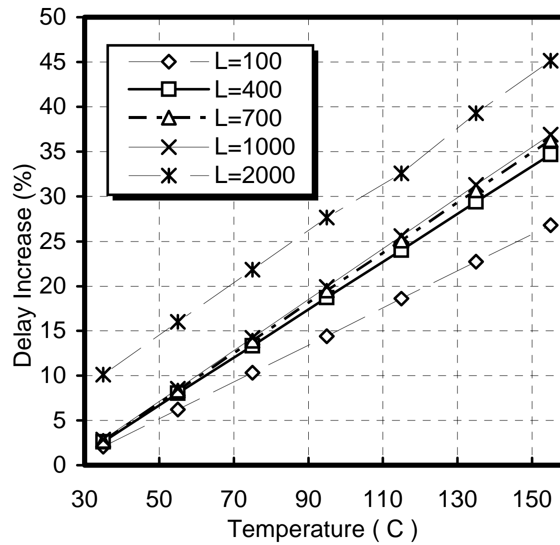

Figure 7. Percentage increase in signal delay with respect to nominal signal delay at room temperature as a function of the interconnect line temperature.

As Figure 7 shows, for each 20-degree increase in temperature there is roughly a 5 to 6 percent increase in the Elmore delay for the long global wires. Although assuming a constant temperature along the interconnect gives an upper bound on the delay increase, we need to estimate and apply the actual variations of temperature along the interconnect lines in (2.22). This is necessary mainly due to the fact that non-uniform interconnect temperature has an unavoidable impact on the wire planning. More specifically, the non-uniform temperature profile along the interconnect line can severely affect the clock

skew and this effect cannot be addressed by simply accounting for a uniform worst-case maximum temperature along the interconnect length [18],[28].

As an example, consider having exponential temperature distributions along the interconnect length. Observing the behavior of the line under exponential thermal profiles is important in the sense that most of the solutions to the interconnect heat transfer equation (2.13) usually have an exponential component. By applying an exponential thermal distribution *T(x)=a.exp(-bx)* to an interconnect and using (2.22), the Elmore delay is as follows:

$$D = D_0 + \frac{a}{b}\rho_0\beta[(c_0L + C_L - \frac{c_0}{b}) + (\frac{c_0}{b} - C_L)e^{-bL}] \quad (2.24)$$

where D_0 is defined by (2.23). For the sake of analysis, consider two different exponential thermal profiles $T_1(x)$ and $T_2(x)$ along an arbitrary interconnect as depicted in Figure 8.

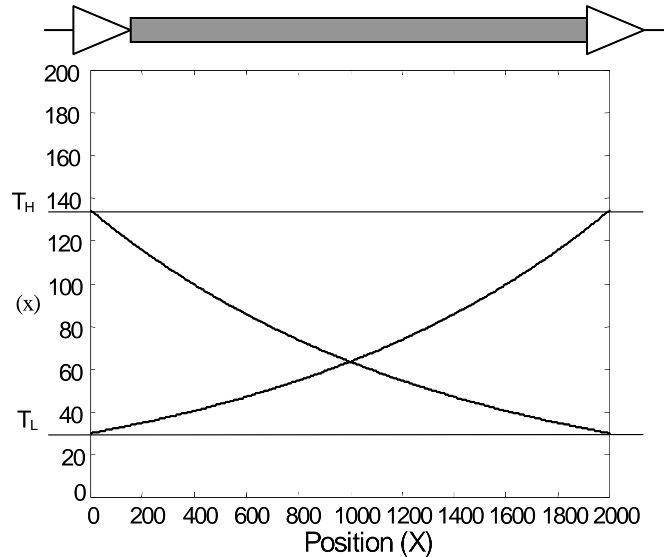

Figure 8. Two different exponentially-distributed thermal profiles along an interconnect line.

Using (2.22), calculation shows that the interconnect Elmore delay is more adversely affected by $T_1(x)$ than by $T_2(x)$, even though the underlying areas for both $T_1(x)$ and $T_2(x)$ in Figure 7 are equal along the length of the line. Figure 9 compares the performance degradation in the presence of $T_1(x)$ and T_2(x) in two different wire lengths, 1000 µm and 2000 µm, having the same electro-thermal characteristics as mentioned before. In both cases the lower-bound temperature is kept constant at 30 °C. By increasing the upper-bound value for these functions, it can be observed that using $T_2(x)$ causes less delay increase than

that caused by using $T_1(x)$. This shows that assuming a constant temperature along the wire (with peak-value) is not accurate enough in planning wire routings and clock-skew analysis, as illustrated later in more detail. The above observation demonstrates that if we have the choice, choosing thermal profile $T_2(x)$ over $T_1(x)$ is preferable. Figure 9 also demonstrates that optimizing thermal profiles is as important as minimizing interconnect length for delay optimization.

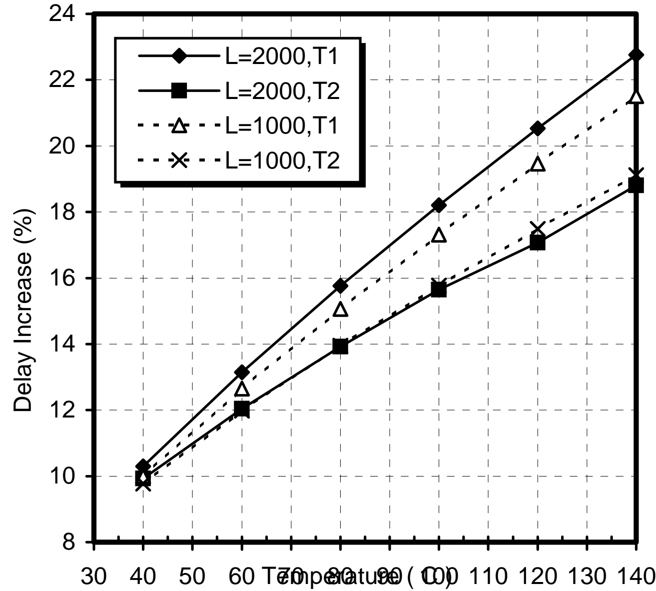

Figure 9. Performance degradation for the $f_1(x)$ and $f_2(x)$ profiles of Figure 2.

It must be noted that the substrate thermal map is strongly dependent on the design, synthesis, floor-planning and placement routines. As a result, analytical modeling of hot spots in the substrate can be a tedious task. However, to approximate a hot spot, one can assume a Gaussian thermal distribution (with constant peak temperature) along the length of a wire with median point μ at a constant peak temperature T_{max} and standard deviation σ as depicted in Figure 10. By applying $T(x) = T_{\max} \cdot \exp(-(x-\mu)^2 / 2\sigma^2)$ to (2.22) we can observe the interconnect performance degradation. The movement of median μ along the length of the line will change the value of the delay degradation, and its effect on performance is also strongly dependent on the value of deviation σ. For the same σ, delay is always better for $\mu=L$ rather than for $\mu=0$ $(0\leq x\leq L)$, which again shows the effectiveness of a gradual increase in the temperature along the line from source to sink. It is obvious that for the same median μ, any increase in the deviation σ will increase the delay. Figure 11 shows the increase in the delay of a wire with length 2000 μm as a function

of different μ's and σ's with T_{max}=120 °C and the same electrical and thermal properties as described above for Figure 7. It can be observed that as μ moves along the line, the location at which the maximum increase in delay occurs is also a function of the deviation σ.

The last two examples illustrate that the delay degradation is strongly dependent on the specific thermal distribution functions. From a resistance point of view, fluctuations of temperature along the line are equivalent to sizing a wire with uniform resistance. In sections with higher temperature, the wire is equivalent to a thinner uniform resistance wire, and in sections with lower temperature the wire acts like a thicker wire with uniform resistance. By recalling the optimization policy in uniform resistance non-uniform wire sizing [29], the best shape for such a line is a decaying exponential from the source of the signal to the destination. Considering the two previous examples of temperature profiles, when the temperature gradually increases from location *0* to *L* the line has a better performance than when there is a gradual decrease in the temperature along the length of the line. Keeping in mind that a gradual *increase* in the line temperature is equivalent to a gradual *decrease* in the size of a uniform resistance line, the results are therefore analogous to optimal uniform resistance non-uniform wire sizing (assuming a constant capacitance).

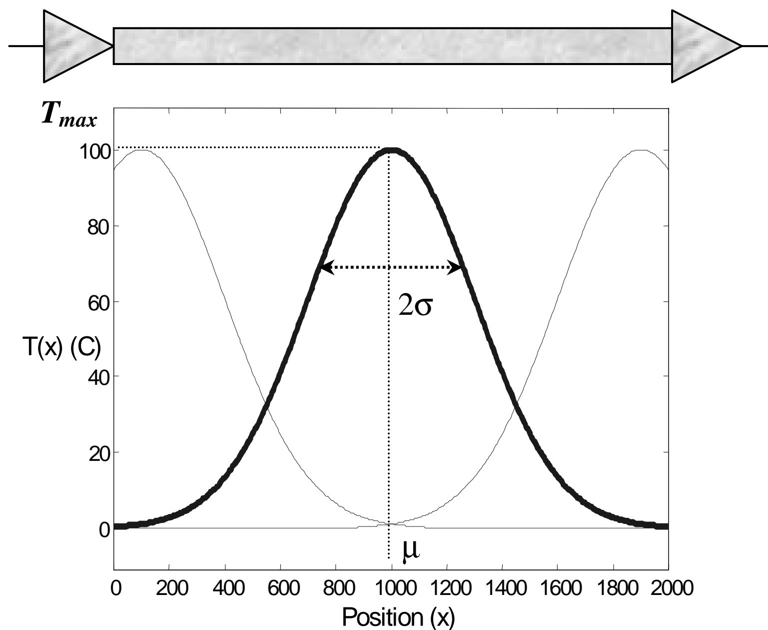

Figure 10. Constant-peak normally-distributed thermal profile with variable median μ and standard deviation σ along an interconnect line.

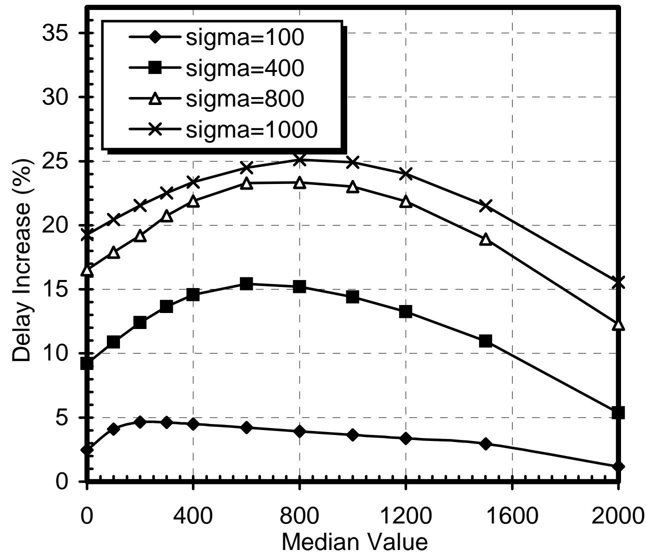

Figure 11. Delay increase as a function of the median value and the standard deviation of a normal temperature distribution.

4 Impact of Non-uniform Interconnect Temperature on Clock Skew

As shown in Section 2, the increase in the Elmore delay can be significant at high temperatures. Moreover, delay variations arising from non-uniform interconnect thermal profiles cannot be accounted for by estimating a worst-case delay based on a uniform maximum temperature along the wires. Consequently, a serious problem may arise, which is the skew fluctuations in a clock signal net. This may in turn degrade the performance of the circuit. Assume a clock net with two fanouts as illustrated in Figure 12. For simplicity assume that both wires 1 and 2 have the same lengths, widths, and electro-thermal characteristics (used in Section 2) and are routed on the same layer.

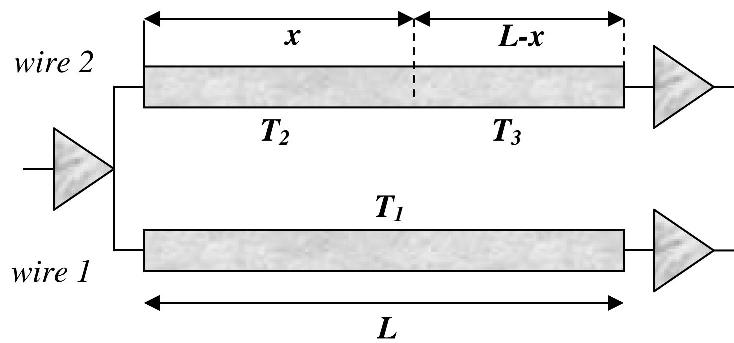

Figure 12. Portion of a clock tree with two fanouts and equal segment lengths.

Assuming different but uniform temperature profiles along both wires, the signal skew can be extracted from Figure 7 by estimating the difference in delay corresponding to the two uniform temperature profiles. A more realistic case arises if one of the wires develops a non-uniform thermal profile along its length due to some underlying thermal gradients over the substrate. In the worst case, we can assume that a section of the line is at one temperature and the rest of the line is at another temperature, as shown in

Figure 12 (for wire 2), with the length x at temperature T_2 and the length $(L-x)$ at temperature T_3. Figure 13 depicts the percentage of the normalized delay increase between wires 1 and 2 as a function of position x in which the thermal gradient occurs at location x, wire 1 is at a uniform temperature of 100 °C, and both the wires are 2000 μm in length. It can be observed that as x approaches zero, the percentage of delay increase reaches its maximum value since the hotter section of the wire (L-x) (which is at T_3) extends over the entire length of the line.

Now assume that with wire 1 remaining at temperature T_1, wire 2 has a certain section of fixed length x where the temperature is lower (or higher) than the rest of the wire. We proceed to study the effect of the magnitude of the gradient between these two sections x and L-x in wire 2 on the normalized delay difference. Assume that temperature T_2 in section x of wire 2 is at uniform temperature of 80 °C while wire 1 is still at uniform temperature of 100 °C. Figure 14 shows that the percentage of normalized delay difference between wires 1 and 2 is a function of the magnitude of the temperature gradient in wire 2. It can also be observed that the magnitude of the thermal gradient is an important factor in the signal skew fluctuations. In this example, due to the specific definition of the thermal gradient, skew becomes zero in a certain location along the length of the wire.

The above analysis shows the importance of considering the effects of the non-uniform interconnect temperature on the clock skew. Due to the high currents driven through the clock wires, clock nets usually exhibit the highest Joule heating among signal nets, and since they span a large area over the die, the probability that they will experience some thermal gradients is much higher than that for the shorter signal nets. As a result, careful consideration of non-uniform temperature profiles is necessary in clock skew estimation along the clock signal net [19].

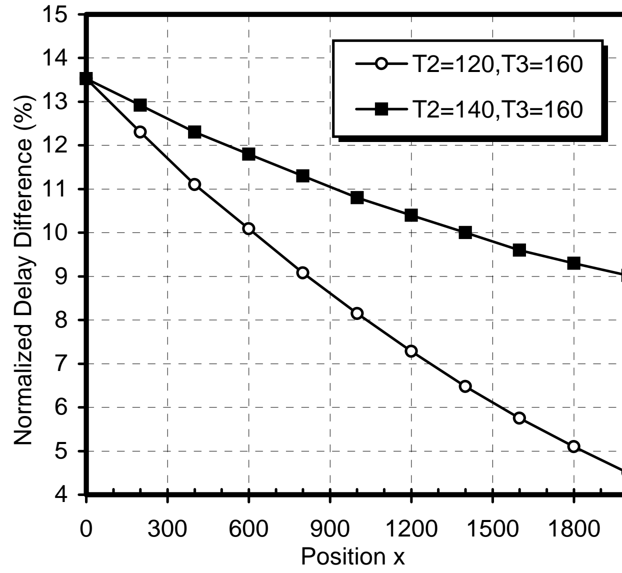

Figure 13. Percentage of normalized delay difference between wires 1 and wire 2 as a function of location parameter *x*.

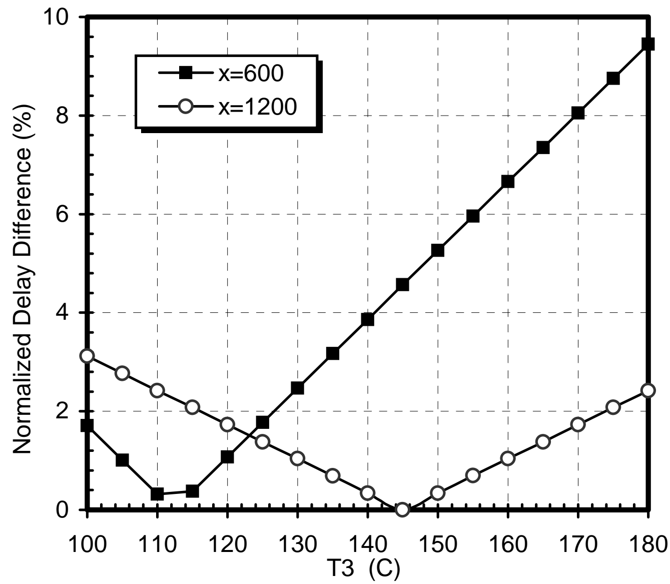

Figure 14. Percentage of normalized delay difference between wires 1 and 2wire as a function of temperature T_3 as shown in Figure 10.

4.1 A Temperature-dependent Zero Skew H-Tree Clock Routing Algorithm

The goal of the clock signal distribution network is to maintain a zero (or near-zero) skew among the sink elements. To ensure zero skew clock distribution, a symmetric H-Tree structure or a bottom-up merging technique can be used. 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 15. 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.

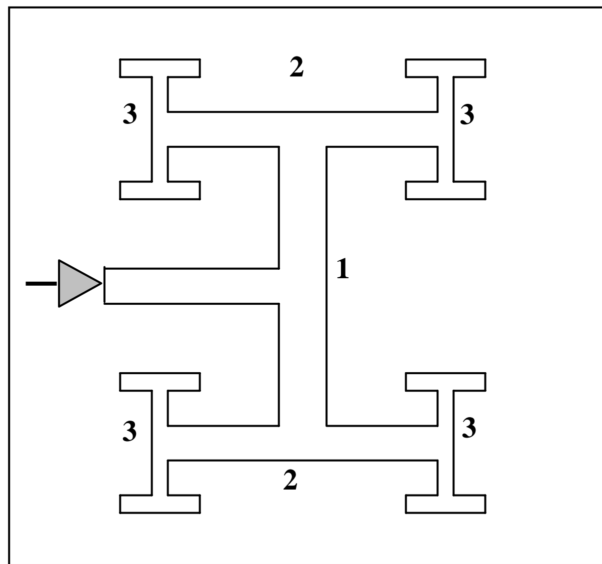

Figure 15. Illustration of a symmetric H-Tree clock distribution net.

The problem arises due to 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 symmetricity of the H-tree cannot guarantee 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, the actual temperature-dependent delay must be considered. Having more than 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*), the locations of the hot spots are in fact quite stable.

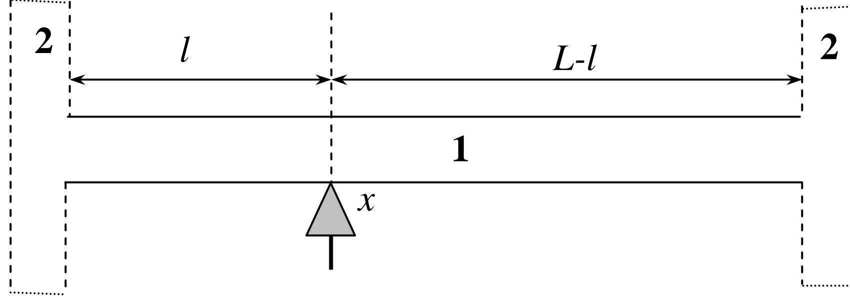

Figure 16. Schematic of minimum-skew clock signal insertion for an interconnect line that is subjected to a non-uniform temperature profile.

Consider the global trunk 1 in the H-Tree depicted in Figure 16. 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 are the same. This will in turn ensure the minimal effect of non-uniform gradients temperature on skew. Assume an interconnect thermal profile $T(x)$ along the length L of trunk 1 and by using the delay model described in Section 3, 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 (2.22), the optimum length l^* for ensuring zero clock skew can be obtained by solving the following equation:

$$\beta \int_0^{l^*} T(x)\,dx + l^* - A = 0 \qquad (5.1)$$

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

$$A = \frac{1}{Lc_0 + C_L}\left(\frac{L^2 c_0}{2} + LC_L + \beta(Lc_0 + C_L)\int_0^L T(x)\,dx - c_0\beta\int_0^L xT(x)\,dx\right) \qquad (5.2)$$

Given circuit parameters L, C_L, c_0, β and $T(x)$, we can easily compute constant A and solve (5.1) to obtain the optimum position for the clock signal connection to the net segment. From (5.1) and (5.2), 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 (5.1), 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) $L/2$.

4.2 Experimental Results

We now examine the behavior of temperature-dependent clock skew for a 2000 μm line with identical electro-thermal characteristics as those in Section 3, 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 via's/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 the fourth column represents the ratio of the clock skew when $l=L/2$ over the delay from the driver to any endpoint of the interconnect line 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. Because 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, changes 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 (5.1) and (5.2) 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.

Thermal Profile	***Parameters***	***l=l****	***Normalized Skew % L=L/2***
$T(x) = ax + b$	T_H=170, T_L=90	1042	5.42
$a = \frac{T_H - T_L}{L}$	T_H=170, T_L=110	1032	3.98
	T_H=170, T_L=130	1021	2.65
$b = T_L$	T_H=170, T_L=150	1012	1.29
$T(x) = a \cdot e^{-bx}$	T_H=170, T_L=90	957.5	5.24
$a = T_H$	T_H=170, T_L=110	968.66	3.63
$b = \frac{1}{L}\ln(\frac{T_H}{T_L})$	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

Table 1. Comparison among different thermal profiles and their impacts on the clock skew.

5 Conclusions

It was shown that non-uniform temperature distributions along global wires in high-performance ICs can have significant implications for interconnect performance. A detailed analysis of the impact of non-uniform temperature distributions on the interconnect performance was presented using a new distributed *RC* delay model that incorporates non-uniform interconnect temperature dependency. The model was applied to analyze a wide variety of interconnect layouts and temperature profiles. Analytical models for accurate interconnect temperature distributions arising from non-uniform substrate temperature profiles were derived using fundamental heat diffusion equations. It was shown that the clock skew can be significantly impacted by the interconnect temperature non-uniformities. These studies suggest that incorporation of thermal analysis is necessary in performing various design optimization steps in high performance ICs.

6 References

[1] V. De and S. Borkar, "Technology and design challenges for low power and high performance," *in Proc. Int'l Symp. on Low Power Electronics and Design*, 1999, pp. 163 –168.

[2] P.P. Gelsinger, “Microprocessors for the new millennium: Challenges, opportunities, and new fronties,” *in Proc. Int’l Solid-State Circuits Conference*, 2000, pp. 22-25.

[3] Y-K Cheng, P. Raha, C-C Teng, E. Rosenbaum, and S. Kang, “ILLIADS-T: an electrothermal timing simulator for temperature-sensitive reliability diagnosis of CMOS VLSI chips,” *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 8, pp.668-681, 1998.

[4] K. Banerjee, A. Mehrotra, A. Sangiovanni-Vincentelli, and C. Hu, “On thermal effects in deep sub-micron VLSI interconnects,” in *Proc. Design Automation Conference,* 1999, pp. 885.

[5] M.T. Bohr, “Interconnect scaling- the real limiter to high performance ULSI,” *in Proc. Int’l Electron Device Meeting*, 1995, pp. 241-244.

[6] K. Banerjee and A. Mehrotra, "Analysis of On-Chip Inductance Effects for Distributed RLC Interconnects," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems,* vol. 21, no. 8, pp. 904-915, 2002.

[7] S. Rzepka, K. Banerjee, E. Meusel, and C. Hu, “Characterization of self-heating in advanced VLSI interconnect lines based on thermal finite element simulation,” *IEEE Trans. on Components, Packaging and Manufacturing Technology-A*, vol. 21, no. 3, pp. 406-411, 1998.

[8] J.R. Black, “Electromigration- A brief survey and some recent results,” *IEEE Trans. on Electron Devices*, vol. ed-16, pp.338- 347, 1969.

[9] Y-K Cheng *et al*, “iCET: A complete chip-level thermal reliability diagnosis tool for CMOS VLSI chips,” in *Proc. Design Automation Conference*, 1996, pp.548-551.

[10] J. Tao, J.F. Chen, N.W. Cheung, C. Hu, “Electromigration design rules for bi-directional current”, *in Proc. International Reliability Physics Symposium*, 1996, pp.180-187.

[11] S. Im and K. Banerjee, “Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs,” *in Proc. Int’l Electron Device Meeting* , 2000, pp. 727-730.

[12] Y. Cheng, C. Tsai, C. Teng, and S. Kang, *Electrothermal analysis of VLSI systems*, Kluwer Academic Publishers, 2000.

[13] Z. Yu, D. Yergeau, R.W. Dutton, S. Nakagawa, N. Chang, S. Lin, and W. Xie, “Full chip thermal simulation,” *in Proc. Int’l Symposium on Quality Electronic Design,* 2000, pp.145-149.

[14] P.E. Gronowski, W.J. Bowhill, R.P. Preston, M.K. Gowan, and R.L. Allmon, “High performance microprocessor design,” *IEEE Journal of Solid-State Circuits*, pp. 676-686, 1998.

[15] Q. Wu, Q. Qiu, and M. Pedram, “Dynamic power management of complex systems using generalized stochastic Petri nets,” *in Proc. Design Automation Conference*, 2000, pp. 352-356.

[16] C.H. Tsai and S.M. Kang, “Cell-level placement for improving substrate thermal distribution,” *IEEE Trans. on Computer Aided Design*, vol 19, no 2, pp 253-265, 2000.

[17] K. Banerjee, M. Pedram, and A.H. Ajami, "Analysis and optimization of thermal issues in high performance VLSI," *in Proc. of Int'l Symposium on Physical Design,* 2001, pp. 230-237.

[18] 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," in Proc. Design Automation Conference, 2001, pp. 567-572.

[19] A.H. Ajami, M. Pedram, and K. Banerjee, "Effects of non-uniform substrate temperature on the clock signal integrity in high performance designs," *in Proc. Custom Integrated Circuits Conference*, 2001, pp. 233-236.

[20] A.J. Chapman, *Fundamentals of heat transfer*, 4th ed., New York, Mcmillan, 1984.

[21] H.A. Schafft, "Thermal analysis of electromigration test structures," *IEEE Trans. on Electron Device*, vol. ed-34, no.3, 1987, pp.664-672.

[22] A.A. Bilotti, "Static temperature distribution in IC chips with isothermal heat sources," *IEEE Trans. on Electron Device*, ed-21, no. 3, pp.217-226, 1974.

[23] R.V. Andrews, "Solving conductive heat transfer problems with electrical-analogue shape factors," *in Chemical Engineering Progress*, vol. 51, no. 2, pp. 67-71, 1955.

[24] D. Chen, E. Li, E. Rosenbaum, and S.M. Kang, "Interconnect thermal modeling for accurate simulation of circuit timing and reliability," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, no. 2, pp.197-205, 2000.

[25] M. Igeta, K. Banerjee, G. Wu, C. Hu, and A. Majumdar, "Thermal characteristics of submicron via's studied by scanning Joule expansion microscopy," *IEEE Electron Device Letters*, vol.21, no.5, pp.224-226, 2000.

[26] C.H. Tsai and S.M. Kang, "Fast temperature calculation for transient electrothermal simulation by mixed frequency/time domain thermal model reduction," *in Proc. Design Automation Conference*, 2000, pp. 750-755.

[27] International technology roadmap for semiconductors (ITRS).

[28] A.H. Ajami, K. Banerjee, and M. Pedram, "Non-uniform chip-temperature dependent signal integrity, " *in Proc. VLSI Symposium on Technology*, 2001, pp. 145-146.

[29] C-P. Chen, Y-P. Chen, and D.F. Wong, "Optimal wire-sizing formula under the Elmore delay model," *in Proc. Design Automated Conference*, 1996, pp.487-490.