Analysis of Substrate Thermal Gradient Effects on Optimal Buffer Insertion

Amir H. Ajami, Kaustav Banerjee*, and Massoud Pedram

Dept. of EE-Systems, Univ. of Southern California, Los Angeles, CA 90089, {aajami, massoud}@zugros.usc.edu

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

Abstract. *This paper studies the effects of the substrate thermal gradients on the buffer insertion techniques. Using a non-uniform temperature-dependent distributed RC interconnect delay model, the buffer insertion problem is analyzed and design guidelines are provided to ensure the near-optimality of the signal performance in the presence of the thermal gradients. In addition, the effect of temperature-dependent driver resistance on the buffer insertion is studied. Experimental results show that neglecting thermal gradients in the substrate and the interconnect lines can result in non-optimal solutions when using standard buffer insertion techniques and that these effects intensify with technology scaling.*

1 Introduction

Aggressive VLSI technology scaling and ever increasing demand for high performance ULSI circuits has resulted in higher current densities in the interconnect lines and higher power dissipation in the substrate [1]. This increase in the power dissipation and the current densities increases the die and the interconnect temperatures. As a result, management of thermal effects is rapidly becoming one of the most challenging efforts in high performance chip design [2]. Furthermore, different activities and sleep modes of the functional blocks in high performance chips cause significant temperature gradients on the substrate. In [3] it has been reported that thermal gradients of 40 °C exist in a high-performance microprocessor design. Low power design techniques such as dynamic power management [4] and clock gating can result in such thermal gradients. With circuits moving toward GHz frequencies it is expected that the magnitude of thermal gradients in the substrate would further increase. In addition, as the minimum feature size shrinks down, the top most metal layers that carry the global signals get closer to the substrate [5]. As a result, the effect of the non-uniform substrate temperature on the interconnect thermal profile becomes more critical. Temperature has an important effect on the circuit performance and reliability [6], [7]. Neglecting thermal gradients in the substrate (and consequently in the interconnects) can introduce major errors in the signal delay calculations [8].

Buffer insertion is an effective technique to reduce the interconnect delay. Some earlier works simultaneously build the fanout tree and insert the buffers [9]. However, most of the buffer insertion techniques start with a fixed tree topology for the fanout net and insert buffers into the tree topology later. Reference [10] finds the optimal delay in a fanout tree by permitting only one non-inverting buffer to be inserted in each segment of the tree. Reference [11] and [12] describe wire segmentation algorithms that allow buffer insertion in a pin-to-pin wire segment of an *RC* fanout tree.

This paper studies the wire segmentation algorithm by considering the influence of the substrate thermal gradients on the performance of the global interconnects and the transistor switching speed. By using a distributed *RC* temperature-dependent delay model, it is shown that in the presence of the substrate temperature gradients the techniques provided in [11] and [12] become non-optimal. To have the maximum efficiency through the buffer insertion, buffers must be sized and placed very carefully along the interconnect line. More precisely [11] and [12] propose equal distances between adjacent buffers in the inserted buffer chain in order to minimize the signal delay. For a non-uniform interconnect thermal profile, we show that the distances between the adjacent buffers do not remain equal and that they are strongly dependent on the thermal profile of the underlying substrate.

A systematic way of calculating the temperature profile of interconnects is described in Section 2. In Section 3 the non-uniform temperature dependent signal delay model is summarized. In Section 4 the effects of substrate temperature on the transistor driver resistance is studied. Section 5 examines the optimality of delay through buffer insertion and by using the non-uniform temperature dependent delay model; new guidelines are proposed to ensure near-optimality of the signal delay. Experimental results including the effect of technology scaling are discussed in Section 6. A brief discussion on the effect of cell driving output resistance on the overall signal performance is presented in Section 7. Finally, concluding remarks are made in Section 8.

2 Non-uniform Interconnect Thermal Profile

The main sources of temperature generation in the chip are the switching activities of the cells in the substrate and the self-heating of the interconnect lines due to the current passing through them. In a high performance design, the substrate temperature can reach up to 120 °C. Self-heating can contribute further to the overall temperature of an interconnect line [5]. Due to the presence of many heat-generating sources in the substrate and the complicated boundary conditions, finding an analytical solution for the heat diffusion equation is non-trivial. Much of the research work has been focused on obtaining a solution by using numerical techniques [13].

By solving the heat diffusion equation and using appropriate boundary conditions in the *3-D* space, the heat flow in an interconnect line can be obtained. In the steady state, with the assumption that the four sidewalls and the top surface of the chip are thermally isolated (which is generally valid), the heat diffusion equation can be reduced to a *1-D* form as follows [14]:

$$\frac{d^2T}{dx^2} = -\frac{Q}{k_m} \qquad (1)$$

where Q is the effective 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 line 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 line 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 the power generation due to the RMS current (I_{rms}) and the rate of the heat loss due to heat transfer between the interconnect and the substrate through insulator. On the other hand, resistance of the interconnect has a linear relationship with its temperature and can be written as follows:

$$r(x) = r_0(1 + \beta \cdot T(x)) \qquad (2)$$

where r_0 is the resistance per unit length at the reference temperature and β is the temperature coefficient of resistance (1/°C). As a result, the heat flow equation (1) in an interconnect can be restated as follows [15], [16]:

$$\frac{d^2T_{line}(x)}{dx^2} = \lambda^2 T_{line}(x) - \lambda^2 T_{ref}(x) - \theta \qquad (3)$$

where λ and θ are constants given as follows:

$$\lambda^2 = \frac{1}{k_m}\left(\frac{k_{ins}}{t_m \cdot t_{ins}} - \frac{I_{rms}{}^2 \cdot \rho \cdot \beta}{w^2 t^2}\right) \qquad (4)$$

$$\theta = \frac{I_{rms}{}^2 \cdot \rho}{w^2 \cdot t_m{}^2 \cdot k_m} \qquad (5)$$

T_{line} is the interconnect temperature as a function of position along the length of the interconnect line and T_{ref} is the underlying substrate temperature. In order to have a unique solution for (3), two boundary conditions must be provided. Equation (3) shows the importance of the substrate temperature profile T_{ref} in determining the interconnect temperature. When considering very short local wires, T_{ref} is usually assumed to be a constant. For long global interconnects, this is not 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 is created by the so-called hot spots in the substrate. As a result, determining the substrate thermal profile is crucial to the thermal analysis of the interconnect lines. From (4) and (5) it can also be seen that the thermal profile along an interconnect line is strongly dependent on the thickness of the underlying insulator t_{ins}. For interconnects assigned to higher metal layers, the thermal resistances between these lines and the substrate are larger than those for local interconnects. Therefore, the higher metal layers experience higher temperatures in comparison to the lower metal layers [7].

3 Temperature-dependent Signal Delay Model

Consider an interconnect line with length L and uniform width w that is driven by a driver with on-resistance R_d and parasitic output capacitance C_p and terminated by a load with capacitance C_L (Fig. 1).

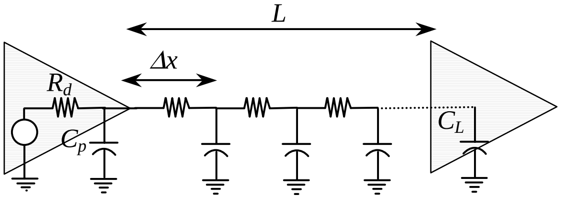

Fig. 1. A distributed RC interconnect line driven by resistance R_d and terminated at load C_L.

The line is partitioned into n equal segments, each with length Δx. Using a distributed RC Elmore delay model and assuming that the number of parts n goes toward infinity, the delay D of a signal passing through the line can be written as follows:

$$D = R_d \cdot (C_p + C_L + \int_0^L c_0(x)\,dx) + \int_0^L r(x) \cdot (\int_x^L c_0(\tau)d\tau + C_L)dx \quad (6)$$

Assume that capacitance per unit length (c_0) does not change with temperature variations along the interconnect length (which is a valid assumption). Also assume that the temperature distribution inside the driver is uniform under the steady-state condition (the R_d will be constant at the chosen operating temperature for the transistor). By using (2), we can rewrite (6) as follows:

$$D = D_0 + (c_0 L + C_L)\, r_0 \beta \int_0^L T(x)dx - c_0 r_0 \beta \int_0^L xT(x)dx \qquad (7)$$

where:

$$D_0 = R_d (C_p + C_L + c_0 L) + (c_0\, r_0 \frac{L^2}{2} + r_0 L C_L) \qquad (8)$$

D_0 is the Elmore delay (at reference temperature) 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 ambient 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, C_p=0 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) [7]. In this calculation we used a uniform temperature profile along the interconnect line which is the worst-case scenario for delay degradation.

It has been shown that the direction of thermal gradients is also an important factor in determining the overall performance of the interconnect line [7]. As a result, the assumption of a constant temperature along the wire (with peak-value) can introduce a large error in planning wire routings. In general, having a gradually increasing thermal gradient results in a better performance than having the same thermal profile in the opposite direction along the length of the interconnect line [7]. 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 have a better delay than that with an increasing sizing profile.

4 Temperature-dependent Driver Resistance

In addition to the dependency of the line resistance to interconnect temperature profile, some CMOS device parameters are also dependent on the substrate temperature including the threshold voltage (V_T), mobility (μ) and energy bandgap of silicon (E_g). It can be shown that thermally dependent variations of mobility and energy bandgap are usually small and not comparable to the threshold voltage variations. For simplicity, assume that the major parameter affected by temperature is the threshold voltage. The first order approximation of the rate of threshold voltage variation as a result of the thermal gradients can be written as follows [17]:

$$\frac{\partial V_T}{\partial T} = \frac{(E_g / q) + V_T}{T} \qquad (9)$$

where T is the device temperature and q is the charge of electron. For silicon the E_g/q is equal to 1.12 Volts. The variation of threshold voltage directly affects the current drawn from the power source and the transistor switching performance. Note that C_L and C_p shown in Fig. 1 will not change with temperature variations. In its simple form, the device driver resistance can be written as follows [18]:

$$R_d = \frac{L_{eff} / w}{\mu C_{ox}(V_{DD} - V_T)} \qquad (10)$$

where L_{eff} is effective channel length, w is the channel width, μ is the mobility, C_{ox} is the gate oxide capacitance and V_{DD} is the power supply voltage. From (9) and (10), it is obvious that threshold voltage variation would causes R_d variations and the rate of driving resistance change due to thermal gradients can be written as follows:

$$\frac{\Delta R_d}{R_d} = \frac{(E_g / q) + V_T}{V_{DD} - V_T} \cdot \frac{\Delta T}{T} \qquad (11)$$

It can be seen that the rate of driver resistance variations due to the temperature fluctuations is strongly dependent on the power supply voltage and threshold voltage (and both of them are technology dependent).

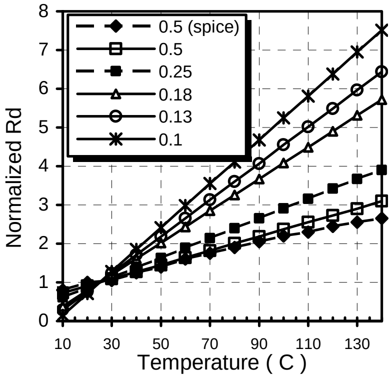

Fig. 2. Normalized driver resistance as a function of device temperature for different technology feature sizes.

Fig. 2 shows the normalized driver resistance as a function of device temperature for different technology nodes based on ITRS specifications [19] (with the assumption of having unit driver resistance at T=25 °C). The SPICE 0.5 µm data has been extracted from [20]. Note that the driver resistance at 25 °C is the base of the normalization (although its actual value depends on the technology feature size).

5 Effects of the Non-uniform Substrate Temperature on Buffer Insertion Techniques

The goal in buffer insertion is to find the number, size and exact location of the inserted buffers along the length of the line, such that the delay is minimized. In [12] it was shown that in a given technology for a buffer chain, there is a critical length between each two buffers, which results in minimization of the delay between the first and the last buffer. In that work, it was assumed that the source, the sink and the buffers have the same size, which results in the same output driver resistance R_d and gate input capacitance C_L for all of them. It was shown that the critical length and the optimal size of the inserted buffers are dependent on the process technology and the interconnect layer assignment and are not dependent on the driver specifications or the number of inserted buffers. The critical length (l_{opt}) and optimal buffer size (s_{opt}) are as follows:

$$l_{opt} = \sqrt{\frac{r_0 c_0 (1+\frac{c_p}{c})}{rc}} \qquad s_{opt} = \sqrt{\frac{r_0 c}{r c_0}} \tag{12}$$

where r_0, c_0, r, c and c_p are the minimum size transistor output resistance (KΩ), minimum size transistor input capacitance (pf), unit length interconnect resistance (KΩ/µm), unit length interconnect capacitance (pf/µm) and parasitic output capacitance (pf) for minimum sized transistor, respectively. In [12] it was assumed that the interconnect is a homogenous line so the r and c are uniform along the length of the line. Having an interconnect with length L, the authors in [11] have shown the optimal number of buffers k with size b which minimizes the delay of the line can be written as follows:

$$k = \lfloor -0.5 + \sqrt{1+\frac{2rcL^2}{R_d(C_L+C_p)}} \rfloor \tag{13}$$

where R_d, C_L and C_p are the buffer output resistance, input capacitance and junction capacitance, respectively. In order to have maximum delay reduction, the buffers must be sized by s_{opt} as stated in (12) (R_d=r_0/s_{opt} , C_L=c_0.s_{opt}, , C_p=c_p.s_{opt}). Moreover, it has been shown that the optimal spacing of the buffers is at equal increments of L/(k+1) on the interconnect line [11]. In this scenario, the propagation delay constant for each segment between two adjacent buffers will be the same for all the k+1 segments as illustrated in Fig. 3.

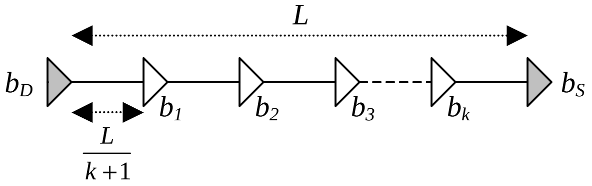

Fig. 3. Structure of standard buffer insertion in a uniform line with equal segmentation.

In Fig. 3 the buffers b_D and b_S are the driver and the sink of the interconnect, respectively, and they are assumed to be at the same size as the buffers b_i (i=1, 2, 3,…, k). However, employing driver and sink with different sizes than the buffers can be easily addressed in (12) and (13) as was shown in [11]. In that case, the distance between the driver (b_D) and the first buffer (b_1) and the distance between the last buffer (b_k) and the sink (b_S) are not equal to the distances between any two adjacent buffers. However, the driver and the sink can always have the same size as the inserted buffers by cascading up or down from some suitable buffers in the library. For simplicity, we assume that the size of the driver, the sink and the inserted buffers are all the same through the rest of this work.

As seen from (2), the interconnect resistance is strongly dependent on the line temperature and any existing gradient in the substrate temperature will affect the signal propagation delay. This suggests that the proposed technique of an equally segmented interconnect does not result in an optimal delay reduction in the presence of a non-uniform substrate thermal profile. Having an interconnect from the source b_D to a single sink b_S, the goal is to find locations of k buffers to be inserted along the length of an interconnect in order to minimize the signal propagation delay subject to a non-uniform substrate thermal profile $T_{ref}(x)$ along the length of the interconnect. The capacitance per unit length c is assumed to be constant and the resistance along the length of the line is a linear function of $T_{ref}(x)$ as stated in (2). We assume that all the k buffer sizes are equal to each other and that can be found using (12). Considering both a variable size for each buffer and a variable distance between each two adjacent buffers makes the problem extremely complicated to solve. As a result we just try to find the exact location of each buffer along the length of the interconnect line.

It must be noted that buffer insertion is generally performed after initial floor planning and cell placement at which point an initial thermal map of the substrate is known. For simplicity, it is assumed that inserting new buffers along the length of an interconnect line does not considerably change the power consumption map of the underlying substrate. This is due to the fact that the power dissipation of these newly inserted buffers is a small fraction of the power dissipation due to the switching activities of the densely placed cells in the surrounding areas. Using the grid-based fast thermal simulation for the substrate [13], the *average* power consumption at each grid cell (which has many individual cells in it) contributes to the overall heat generation of that area. As a result, it is assumed that the temperature of each inserted buffer will eventually reach a steady-state value equal to the local substrate temperature.

Based on (11), the cell driver resistance is a linear function of the local substrate temperature and can be written as follows:

$$R_d(x) = R_{d0}(1+\beta_c \cdot T_{ref}(x)) \tag{14}$$

where $R_d(x)$ is the driver resistance profile of the transistors with thermal profile $T_{ref}(x)$, R_{d0} is the cell driver resistance at reference temperature (namely 25 °C) and β_c is the temperature coefficient of driver resistance (1/°C). β_c can be extracted from SPICE simulations at different temperatures or by using (11). Based on Fig. 3 with buffer locations x_1, x_2, x_3,…, x_k as variables, the path delay from source to sink can be written as follows:

$$D=\sum_{i=1}^{k+1}(\int_{x_i}^{x_{i+1}} R(\tau)(c(x_i-\tau)+C_L)d\tau)+\sum_{i=1}^{k+1} R_d(x_{i-1})(cx_i-cx_{i-1}+C_L+C_p) \tag{15}$$

where x_0 and x_{k+1} are constants equal to zero and L respectively. The first term is the interconnect delay while the second term represents the gate delay. Based on the functionality of $T_{ref}(x)$ and dependency of $R(x)$ and $R_d(x)$ on it (as shown in (2) and (14)), path delay (15) may or may not behave as a convex objective function. In case of being convex, the global minimum can be obtained by solving the systems of k partial differential equations (by using the partial derivatives of D with respect to variables x_i's (i = 1, 2, 3, …, k) and setting them to zero). In general, the derivative of D with respect to the position of the ith buffer (1 ≤ i ≤ k) can be written as follows:

$$\frac{\partial D}{\partial x_i}=\int_{x_{i-1}}^{x_i} R(\tau)d\tau+(x_i-x_{i+1})R(x_i)+(x_{i+1}-x_i+\frac{C_L+C_p}{c})\frac{\partial R_d(x_i)}{x_i}-R_d(x_{i-1})-R_d(x_i) \tag{16}$$

In case of having a non-convex optimization problem, solving the k partial differential equations may result in a local minimum. One can use the Quasi-Newton method to approximate the delay objective locally by a quadratic function that can be further minimized near-globally with a total order of convergence of at least two [21].

6 Experimental Results

Now the effects of the non-uniform interconnect temperature profile on the buffer insertion problem are examined. Even though the method presented in section 5 derives the location of the inserted buffers, we still need to find the optimal number of buffers, k. Due to the non-uniform thermal profile of the interconnect, the line resistance per unit length is a function of the position along the length of the line which has a minimum r_{min} and maximum r_{max}. These two values are used in (13) to find out the optimal number of buffers that need to be inserted into a line. Using the extreme values of the line resistance per unit length (r_{min} and r_{max}) may result in different values for k. In that case both values are examined and the minimum delay resulting from using the suitable number of buffer stages for each case is used.

Table 1 shows the parameters used in our experiments for different technology nodes based on ITRS specifications [19]. By using a simple linear function ($ax+b$) of the position x for the substrate temperature $T_{ref}(x)$, the temperature-dependent buffer insertion technique is examined. Note that in reality the substrate thermal profile is dependent on the design, synthesis, floor planning and placement techniques and the temperature profile along the substrate may not be a linear function of x. In this example, it is assumed that a 75 °C thermal gradient between the two ends of the wire is present (from 25 °C to 100 °C). Furthermore, it is also assumed that the left side (b_D in Fig. 3) of the interconnect line is the cooler side at 25 °C.

Parameter	***0.18μm***	***0.13 μm***	***0.1μm***
r (KΩ/m)	36.3	60.1	103.9
c (pf/m)	269	240	154
C_L (ff)	1.9	1.7	1.5
R_d (KΩ)	8	9.5	10
C_p (ff)	4.8	3.5	2.5
l_{opt}(mm)	3.33	2.5	2.22
s_{opt}	174	151	110
V_{DD} (V)	1.8	1.5	1.2

Table 1. Parameters used in generating experimental results for three different technologies based on ITRS specifications and [22].

For illustrative purposes, we consider two cases while optimizing the objective function (15): *(i)* non-uniform driver resistance R_d, uniform interconnect resistance per unit length r and *(ii)* non-uniform driver resistance R_d, non-uniform interconnect resistance per unit length r. Fig. 4 shows the percentage of performance improvement for temperature aware buffer insertion in comparison to the standard buffer insertion techniques using buffers from the library with optimal size and optimal length between each two adjacent buffers defined by (12) for different technology nodes. In Fig. 4, the vertical axis shows the percentage decrease in the signal propagation delay. The graphs labeled with R_d are those related to the analysis where only R_d is considered as a temperature dependent variable, while the graphs labeled with (R_d+r) consider both temperature dependent R_d and r. It can be seen that as the feature size shrinks down, the effects of the substrate thermal non–uniformities on signal performance becomes more critical. It can also be observed that as the interconnect length increases (which results in increased number of inserted buffers) the improvement in the signal delay becomes less than that for the shorter lines and it will eventually saturate to a lower bound.

One notable fact is that the performance improvement in the case of only variable R_d is more than that the case with both variable R_d and variable r. From the interconnect resistance point of view, in order to minimize the signal delay an increasing thermal profile from source to sink requires that the inserted buffers to move to the *right* side to reduce the length of the interconnect between each two adjacent buffers in the areas with higher interconnect temperature. This can be seen in Fig. 5. Fig. 5(a) shows the standard buffer insertion result for one buffer. Fig. 5(b) shows that by considering only a variable interconnect resistance (r) and constant R_d, the location of the inserted buffer must be shifted to the right side to reduce the length of the section with higher temperature. However, from the driver resistance point of view, using an increasing thermal profile from source to sink, forces the inserted buffer to move to the *left* side to reduce the device resistance as much as possible as shown in Fig. 5(c). From Fig. 5(b) and 5(c), it can be observed that the movement of inserted buffers is more severe in the case of having a temperature dependent R_d instead of having a temperature dependent r. This is expected, as the magnitude of R_d is much more than that of the interconnect resistance per unit length. In addition, the device driver resistance dependency on the temperature is much more severe than that for the interconnect resistance. As a result, by considering both variable R_d and variable r, the inserted buffer lies between the computed locations of the buffer for the case of Fig. 5(b) and 5(c) as shown in Fig. 5(d).

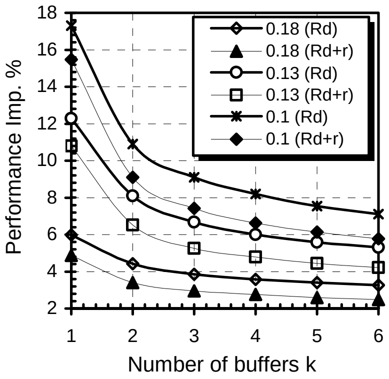

Fig. 4. Delay improvement due to the temperature-aware buffer insertion technique in comparison to the standard buffer insertion technique for different number of buffers in different technologies based on ITRS.

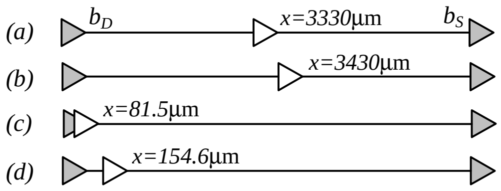

Fig. 5. Location of an inserted buffer in a 6660 μm line (0.18 μm technology): (a) standard technique (b) temperature-aware technique with only variable r (c) temperature-aware technique with only variable R_d (d) temperature-aware technique with both variable R_d and r.

Fig. 6 shows the dependency of delay improvement on the magnitude of thermal gradient between the two ends of the wire for different technologies. It can be observed that as the gradient increases, the standard buffer insertion techniques become less efficient. This shows the importance of having a temperature aware buffer insertion technique knowing the fact that in a high performance design a 40 °C to 50 °C thermal gradient is inevitable and that these gradients tend to increase for the future technologies.

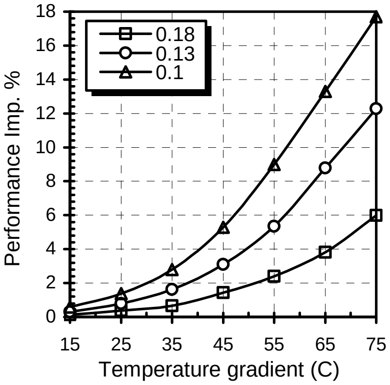

Fig. 6. Delay improvement due to the thermally aware buffer insertion for one buffer as a function of different thermal gradients between the two ends of the line in comparison to the standard buffer insertion techniques for different technologies.

Fig. 7 shows the dependency of the performance improvement as a function of percentage of optimal length (l_{opt}) defined by (12) which shows that thermally aware buffer insertion will be more effective in interconnects with critical lengths less than the optimal length between each two adjacent buffers provided by Table 1. This suggests that the optimal buffer size and total optimal length of the wire connecting the source and the sink can also be adjusted in the presence of non-uniform thermal profiles along the length of the interconnect line.

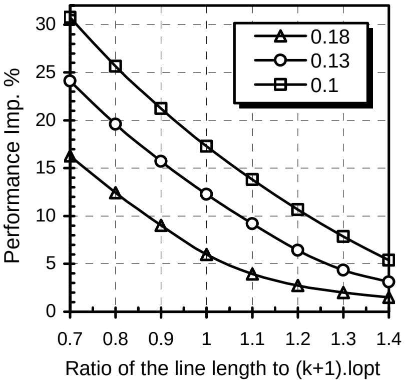

Fig. 7. Delay improvement due to the thermally aware buffer insertion for one buffer as a function of percentage of critical length in comparison to the standard buffer insertion techniques for different technologies.

7 Discussion

It is observed that the variations of driver resistance R_d dramatically affect the optimality of the solution in the buffer insertion problem. As a result, determining sources of such driver resistance variations is a crucial step in formulating the objective functions in different EDA flow steps. Device temperature variation is an important source of R_d variation. Process variations can further contribute to the R_d fluctuations. From (11) , it can be deduced that technology feature size and the power supply voltage are also controlling the driver resistance. Power supply voltage variations are caused mainly by the IR-drop effect. In addition, substrate thermal gradients will affect both the signal interconnect and the power grid resistances. It is expected that these thermal gradients will affect the IR-drop and in the worst-case increase it significantly. As a result, for finding near-optimal solutions, one must take various thermal effects into consideration (knowing the fact that as the technology feature size shrinks down, thermal effects become much more severe).

8 Conclusion

In conclusion, an analysis of the impact of non-uniform substrate temperature distributions on the buffer insertion scheme along global lines has been presented using a distributed RC delay model that incorporates the non-uniform interconnect temperature dependency. It is shown that non-uniform temperature distributions along the substrate in high-performance ICs can have a significant impact on the interconnect performance and cell propagation delay. An analytical model for addressing the effects of substrate temperature non-uniformities on the position of inserted buffers along the interconnect lines has been formulated. It is observed that the delay degradation caused by the effects of temperature on the cell driver on-resistance are much more severe than that caused by the interconnect resistance thermal dependency. It is shown that as VLSI technology scales down, these non-uniform thermal effects will become more severe and must be taken into account in the design methodology to ensure near-optimality of the performance.

As an extension to this work, the effects of thermal gradients on the buffer insertion techniques for trees and multi-port signal nets will be studied. Based on the algorithms proposed in [10] and [11], the thermally aware buffer insertion technique with wire segmentation will be extended to include the insertion of the buffers in a tree network. Also as a future study, the effect of newly inserted buffers on the overall substrate temperature map will be analyzed. In this work it is assumed that those effects are negligible. However this condition can be relaxed such that effects of the inserted buffers on the overall substrate temperature are considered. As a result, an iterative procedure will be employed to calculate the updated average temperature in the vicinity of the inserted buffers, and new thermally aware buffer insertion procedure will take place till convergence.

Acknowledgement

This work was supported in part by the SRC under contract number 98-DJ-606. The authors would like to also acknowledge support from the MARCO Interconnect Focus Center at Stanford University.

9 References

[1] V. Tiwari, D. Sing, S. Rajgopal, G. Mehta, R. Patel, and F. Baez, “ Reducing power in high-performance microprocessors,” *Proc. Design Automation Conf.*, 1998, pp. 732-737.

[2] L. Gwennap, “Power issues may limit future CPUs,” *Microprocessor Report*, 10(10), August 1996.

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

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

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

[6] K. Banerjee A. Mehrotra, A. Sangiovanni-Vincentelli, and C. Hu, “On thermal effects in deep sub-micron VLSI interconnects,” *Proc. Design Automation Conf.*, 1999, pp. 885-891.

[7] 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,” *Proc. Design Automation Conf.*, 2001, pp. 567-572.

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

[9] K.J. Singh and A. Sangiovanni-Vincentelli, “A heuristic algorithm for the fanout problem,” *Proc. Design Automation Conf.*, 1990, pp. 357-360.

[10] L.P.P.P van Ginneken, “Buffer placement in distributed RC-tree networks for minimal Elmore delay,” *Proc. Int. Symp. on Circuits and Systems*, 1990, pp. 865-868.

[11] C. Alpert and A. Devgan, “Wire segmenting for improved buffer insertion,” *Proc. Design Automation Conf.*, 1997, pp. 588-593.

[12] R.H.J.M. Otten and R.K. Brayton, “Planning for performance,” *Proc. Design Automation Conf.*, 1998, pp. 122-127.

[13] Y. Cheng, C. Tsai, C. Teng, and S. Kang, *Electrothermal Analysis of VLSI Systems*, Kluwer Academic Publishers, 2000.

[14] A.J. Chapman, *Fundamentals of Heat Transfer*, Mcmillan Inc., 1984.

[15] H.A. Schafft, “Thermal analysis of electromigration test structures,” *IEEE Trans. on Electron Device*, vol. Ed-34, No.3, pp. 664-672, 1987.

[16] A.A. Bilotti, “Static temperature distribution in IC chips with isothermal heat sources,” *IEEE Trans. on Electron Device*, vol. Ed-21, No. 3, pp. 217-226, 1974.

[17] E.S. Yang, *Microelectronic Devices*, McGraw-Hill Inc., 1988.

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

[19] *The International Technology Roadmap for Semiconductors (ITRS)*, 1999.

[20] R. Kielkowski, *SPICE Practical Device Modeling*, McGraw-Hill Inc., 1995.

[21] D.G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley Inc., 1984.

[22] K. Banerjee and A. Mehrotra, "Accurate analysis of on-chip inductance effects and implications for optimal repeater insertion and technology scaling," *Proc. Symp. on VLSI Circuits*, 2001, pp. 195-198.