Temperature Dependent Wire Delay Estimation in Floorplanning

Winther, Andreas Thor; Liu, Wei; Nannarelli, Alberto; Vrudhula, Sarma

Published in:
Proceedings of NORCHIP 2011

Link to article, DOI:
10.1109/NORCHP.2011.6126741

Publication date:
2011

Temperature Dependent Wire Delay Estimation in Floorplanning

Andreas Thor Winther, Wei Liu*, Alberto Nannarelli and Sarma Vrudhula†

Dept. of Informatics, Technical University of Denmark, Kongens Lyngby, Denmark
*Dept. of Computer Engineering, Politecnico di Torino, Torino, Italy
†Computer Systems Engineering, Arizona State University, Tempe, USA

***Abstract*—Due to large variations in temperature in VLSI circuits and the linear relationship between metal resistance and temperature, the delay through wires of the same length can be different. Traditional thermal aware floorplanning algorithms use wirelength to estimate delay and routability. In this work, we show that using wirelength as the evaluation metric does not always produce a floorplan with the shortest delay. We propose a temperature dependent wire delay estimation method for thermal aware floorplanning algorithms, which takes into account the thermal effect on wire delay. The experiment results show that a shorter delay can be achieved using the proposed method. In addition, we also discuss the congestion and reliability issues as they are closely related to routing and temperature.**

I. Introduction

With technology scaling, the feature sizes of both CMOS devices and wires shrink and designers are able to integrate more and more functionalities into a single chip. Delay in CMOS transistors decreases as the channel length is reduced in each new process. Delay in metal wires, on the other hand, shows different behaviors. For local wires, delay decreases as the distance between the end points becomes smaller with scaling. For global wires, which has to span across the chip, delay increases due to the fact that die size does not shrink but slightly increases in each new process. In fact, delay in global wires has increased steadily with technology scaling over the years and already dominates path delays.

In addition to technology scaling, modeling of global wires is further complicated by thermal effects. Due to the high degree of integration and aggressive power management techniques (clock gating, power gating, etc.), the power consumption in different regions of the chip (e.g. the power density) can vary significantly. The spatially non uniform power consumption within the chip exhibits as thermal gradients, which are temperature differences between different regions.

The high temperature and large thermal gradient in metal layers can affect many aspects of interconnect design, including signal delay, routing congestion and reliability. The propagation delay in metal wires is severely degraded by high temperature as the electrical resistivity in metal increases linearly with temperature. The large within-die thermal gradients result in performance mismatch between wires of the same length but subject to different temperatures. Traditional physical design algorithms such as floorplanning and routing assume resistivity in interconnects is uniform and constant. Consequently, wirelength is used as a metric to estimate signal delay and congestion of interconnects. However, in designs where the substrate has nonuniform thermal profile, the traditional way of estimating wire delay can lead to large errors. This is because wire performance decreases with an increase in temperature and the delay of two wires of the same length are no longer equal.

The thermal effect is more prominent in global wires than in local wires because global wires are routed in layers that are far away from the heat sink, and because global wires are routed for long distance and may develop a large thermal gradient. In recent years, temperature variation induced clock skew in clock distribution network has received a lot of attention. In [1], [2], the authors described design time clock tree synthesis algorithms to modify merging locations against nonuniform substrate thermal profile. While in [3], optimal insertion of tunable delay buffers into clock trees is discussed to adjust at run time the delay of clock distribution paths that are more susceptible to temperature variations. Thermal aware global routing algorithms for improving reliability are also discussed in [4], [5].

On the other side, regarding global signal wires, although extensive work has been done on thermal aware floorplanning, all of them assume electrical resistivity in wires is constant and thermal gradients in the substrate has no impact on wire delay. This assumption is in general invalid and increasingly inaccurate in nanometer high performance designs where large temperature gradients already exist in the substrate.

In this paper, we study the problem of estimating the temperature dependent wire delay during the floorplanning stage. We first illustrate the impact of nonuniform thermal profile on the delay in wires. Then we propose a new way to estimate the wire delay in thermal aware floorplanning algorithms. The proposed algorithm takes the delay, instead of the wirelength, as one of the optimization goals, in this way, mitigating the excessive delay increase caused by high temperature. In addition, we also consider the impact of routing congestion and the reliability of wires, which are important metrics in evaluating floorplans in a realistic setting.

II. Thermal Aware Floorplanning

Floorplanning is the initial stage of physical implementation of VLSI circuits, which to a large extent determines the

quality of the final design. During the floorpalnning stage, the main design tasks include macro block placement, global wire planning and Power/Ground network design. Traditional floorplanning algorithms only optimizes the total area and wirelength. In recent years, as thermal issues become more prominent, the maximum temperature is also added to the cost functions in so called thermal aware floorplanning algorithms [6], [7], [8]. Since the thermal coupling between high power consumption blocks can significantly affect the temperature distribution in the whole chip, thermal aware floorplanning algorithms consider peak temperature in addition to area and wirelength in the evaluation of a floorplan. The estimation of peak temperature usually requires the use of compact thermal models that can compute the temperature profile in a very efficient way [9].

The floorplanning tool proposed in [6], HotFloorplan, is a very representative thermal aware floorplanner. The topology of the floorplan is represented in *Normalized Polish Expression* and the optimization process is implemented as a simulated annealing process. During the annealing, for every candidate floorplan, the algorithm invokes routines in HotSpot [10] to compute the temperature distribution and uses the maximum temperature as one of the metrics in the cost function. The other metrics in the cost function are total area and total wirelength. The connectivity information in HotFloorplan is stored in a two-dimensional connectivity matrix and manhattan distance is used to estimate the wirelength between two endpoints in a wire.

III. Temperature Estimation and Wire Delay Calculation

The high temperature in global wires is caused not only by self heating, but also by heat diffusion from the substrate. According to [11], the temperature within the interconnect for a given substrate temperature can be expressed as:

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

where θ and λ are constants for a chosen metal layer in a specific technology node and depend on the thermal conductivity of metal and insulator, on their geometries, and on the electical parameters of the interconnect (current density and resistivity).

The peak temperature rise is equal to θ/λ^2 for interconnects whose lengths (L) are larger than the heat diffusion length. As the global wires are routed in top metal layers, the distance from the substrate is larger than local wires, and, as a result, the temperature rise in global wires is higher.

The electrical resistance of metal has a linear relationship with its temperature and can be expressed as:

$$R(x) = R_0(1 + \beta \cdot T(x)) \quad (2)$$

where R_0 is the resistance at reference temperature, β is the temperature coefficient ($1/^\circ$C) and $T(x)$ is the temperature profile along the length of the wire. The value of β for copper at room temperature is 3.9E-3, which means for every 10 $^\circ$C increase in temperature, the resistance would increase by 3.9%. According to the distributed RC Elmore delay model [12], signal propagation delay through the interconnect of length L can be written as:

$$D = R_d\left(C_L + \int_0^L c_0(x)dx\right) + \int_0^L r_0(x) \cdot \left(\int_x^L c_0(\tau)d\tau + C_L\right)dx \quad (3)$$

where R_d is the driver cell's ON resistance, $c_0(x)$ and $r_0(x)$ are the capacitance and resistance per unit length at location x and C_L is the load capacitance.

By combining (2) and (3), we can obtain a temperature dependent interconnect delay model:

$$D = D_0 + (c_0 L + C_L) r_0 \beta \int_0^L T(x)dx - c_0 r_0 \beta \int_0^L x \cdot T(x)dx \quad (4)$$

where

$$D_0 = R_d(c_0 L + C_L) + \left(c_0 r_0 \frac{L^2}{2} + r_0 L C_L\right) \quad (5)$$

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

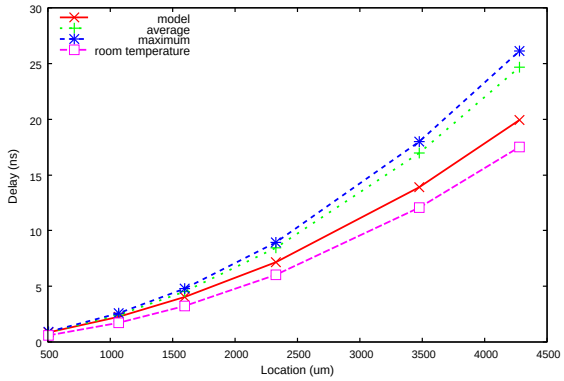

Fig. 1. Delay increase in a global wire using different thermal profiles

In Figure 1, we plot the delay increase in a global wire from one benchmark circuit in a 50 nm process using different thermal profiles. The red curve in solid line is the delay calculated using (4) with a thermal profile extracted from HotFloorplan. For comparison purposes, we also plot the delay increase when uniform temperature profiles are used, namely the maximum, the average and the room temperature. The delay at the end of the wire using the extracted thermal profile is 20 ns while using the maximum and average temperatures can result in errors larger than 25%. This means when estimating global wire delay, we need to take the thermal gradient and not the maximum or average temperature into consideration.

IV. Thermal Aware Wire Planning in Floorplanning

As we have discussed in Section III, the delay of global wires subject to large temperature variations is no longer linearly

proportional to wirelength. To have an accurate estimation of wire delay, the temperature effect including thermal gradients has to be considered. Since the thermal profile on the die is mainly determined by the locations of macro blocks, it is, therefore, possible to perform temperature dependent delay estimation at the floorplanning stage.

Our wire planning method is described in Algorithm 1. Given a floorplan, together with the associated connectivity matrix and the thermal map, we compute the total delay, the maximum congestion and the average reliability of all wires. The layout of each wire is determined by performing L-shape routing between the center of the connecting blocks. Once the physical layout of the wire is known, we record the blocks over which the wire is routed. The temperature profile along the wire and the wire delay are then calculated using (1) and (4). With thermal effects taken in account, the two paths in the bounding box of two end points of a wire can have different delay although their length are the same. In our algorithm, we choose the path with a shorter delay, which is different from HotFloorplan where the two paths are considered as identical. The temperature profile is also used to evaluate the wire reliability in terms of Mean Time To Failure (MTTF). In addition, a congestion map made up of a two dimensional matrix is updated with the route of the wire to evaluate the routability of the floorplan. The congestion map is useful because in our algorithm more wires are likely to be routed in regions with a low temperature, potentially causing routing congestion. We now describe the congestion map and the reliability metrics used in the algorithm.

Algorithm 1 Wire planning in floorplanning

INPUT: Floorplan description
INPUT: Connectivity matrix
INPUT: Thermal map
for all wires in the connectivity matrix **do**
  Perform L-Shape routing
  Extract the thermal profile along the wire
  Calculate wire delay using (4)
  Calculate wire reliability
  Update congestion map
end for
OUTPUT: Delay, Congestion and Reliability of Wires

A. Congestion Map

Our congestion map is made up of a two dimensional matrix, which divides the floorplan into a congestion grid. The size of the routing cell in the grid is 80 μm $\times$ 80 μm. During the routing of wires, the value of a cell is incremented if a wire passes through the cell. A higher value means more wires passes through that cell.

An example congestion map is shown in Figure 2, where the congestion cells are plotted in colors over the floorplan. Cells in red color indicate congestion hotspots. The maximum value in the congestion matrix can be used to evaluate the routability of the design.

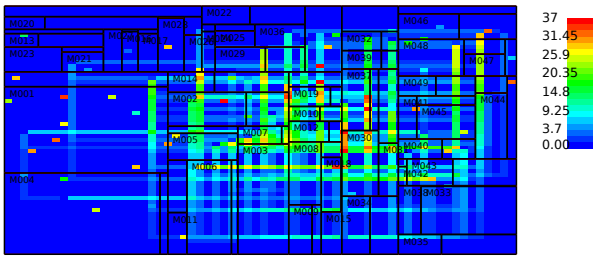

Fig. 2. Example congestion map

B. Reliability Metrics

The reliability of wires is usually measured by MTTF, which is an estimation of the average time before the wire fails due to electromigration. The MTTF can be modeled using Black's equation [13],

$$MTTF = \frac{A}{J^2} e^{\frac{Ea}{KT}} \tag{6}$$

where T is the wire temperature, J is the current density, A is related to the wire cross-sectional area and Ea and K are the activation energy and Boltzmann constants respectively.

To capture the exponential relationship between the wire reliability and temperature, we adopt the relative reliability metric used in [5], which computes the ratio of the MTTF of a wire under room temperature T_{rm} to the MTTF under temperature T_e,

$$r(e) = \frac{MTTF^{rm}}{MTTF^{e}} = exp\left[\frac{Ea}{K}\left(\frac{1}{T_{rm}} - \frac{1}{T_e}\right)\right] \tag{7}$$

In Algorithm 1, we compute the relative reliability of each wire and use the average value to evaluate the reliability of the floorplan.

V. Experimental Results

We implemented the wire planning algorithm described in the previous section in HotFloorplan and to speed up the lengthy simulation time of HotFloorplan we adopt the fast simulated annealing (FastSA) approach proposed in [14]. We used the MCNC macro block benchmark circuits with 50 nm process parameters to evaluate our proposed wire planning methods. To compare the proposed methods against the original HotFloorplan implementation, we run the floorplanning algorithm on each benchmark circuit using different cost functions.

The experiment results are summarized in Table I, where *Cmax* and *Ravg* in the last two columns are the normalized maximum congestion value and the normalized average reliability of all wires respectively. The three cost functions we used are:

- **CF1** is the original function in HotFloorplan defined as $\lambda_A \times Area + \lambda_W \times Wirelength + \lambda_T \times Tmax$
- **CF2** replaces Wirelength with WireDelay in CF1
- **CF3** adds the maximum congestion and the average reliability to CF2

In the experiment results, **CF1** and **CF2** are of the most interest. Since the weight of each metric (λ) in the cost

TABLE I
EXPERIMENTAL RESULTS ON MCNC BENCHMARKS USING DIFFERENT COST FUNCTIONS

Benchmark	*# Blocks*	*Cost Function*	*Tmax (K)*	*Area (mm^2)*	*Wirelength (m)*	*Tot. Delay (μs)*	*Cmax*	*Ravg*
ami49	49	CF1	386.7	39.9	5.518	5.55	1.00	1.00
		CF2	389.4	41.7	5.456	4.94	1.43	0.76
		CF3	389.5	43.0	5.751	5.18	0.71	0.82
ami33	33	CF1	394.0	1.3	8.438	6.02	1.00	1.00
		CF2	396.4	1.3	8.172	5.80	1.01	1.01
		CF3	385.1	1.5	9.096	6.42	0.49	0.67
apte	9	CF1	378.3	48.2	2.834	4.46	1.00	1.00
		CF2	379.4	48.2	2.723	4.20	1.00	0.95
		CF3	378.7	48.4	3.146	5.07	0.72	0.95
hp	11	CF1	360.5	9.5	1.732	2.57	1.00	1.00
		CF2	358.4	10.9	2.026	2.50	1.04	0.85
		CF3	358.0	10.9	1.786	2.01	0.67	0.69
xerox	10	CF1	367.9	20.9	20.315	25.03	1.00	1.00
		CF2	368.7	21.0	20.724	19.98	0.91	0.84
		CF3	369.5	21.3	23.612	24.98	0.82	0.88

functions can significantly alter the results, we list **CF3** only for comparison purposes. The λs we chose in **CF1** and **CF2** give order of optimization priorities from high to low as: area, peak temperature and wire performance. *Cmax* and *Ravg* are normalized to **CF1**, which give a comparison against the results obtained from the original HotFloorplan.

It can be seen in Table I that in all cases using delay (**CF2**) instead of wirelength (**CF1**) to measure the wire performance always produces floorplans with a smaller total delay. The amount of reduction in total delay can be significant, 11% decrease in *ami49* and 20% decrease in *xerox*. The cost is usually a slight increase in area and wirelength. For example in *ami49* where the number of blocks is much larger than the others, the area in **CF2** is increased by 4.6% from **CF1**.

On the other hand, due to the thermal awareness in the algorithm, there is always an improvement in reliability of wires in **CF2** and **CF3**. Note that reliability is relative to room temperature and according to (7) a smaller value indicates a longer MTTF. As mentioned in Section IV, avoiding thermal hotspots during routing might cause congestion in the cool area. From Table I we can see that, in the worst case, as in *ami49*, the maximum congestion increased by 43% can potentially cause a routing problem. When congestion is also taken into consideration (as in **CF3**), the routing can be made much easier at the cost of an increase in area and wirelength.

VI. CONCLUSIONS

In this paper, we proposed a wire delay estimation method at the floorplanning stage which takes into account the performance degradation in wires due to thermal effect. The method is implemented in HotFloorplan and evaluated using the MCNC benchmarks with congestion and wire reliability considered. The experiment results show that in the presence of thermal gradients shorter wirelength does not always produce shorter delay. The proposed method, on the other hand, can achieve a better total delay and wire reliability at the cost of an increase in area and wirelength.

REFERENCES

[1] C. Liu, J. Su, and Y. Shi, "Temperature-aware clock tree synthesis considering spatiotemporal hot spot correlations," *Proc. of 26th IEEE International Conference on Computer Design*, pp. 107–113, Oct. 2008.

[2] M. Cho, S. Ahmedtt, and D. Pan, "TACO: temperature aware clock-tree optimization," Nov. 2005, pp. 582–587.

[3] A. Chakraborty, K. Duraisami, A. Sathanur, P. Sithambaram, L. Benini, A. Macii, E. Macii, and M. Poncino, "Dynamic Thermal Clock Skew Compensation Using Tunable Delay Buffers," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, no. 6, pp. 639–649, June 2008.

[4] A. Gupta, N. Dutt, F. Kurdahi, K. Khouri, and M. Abadir, "Thermal Aware Global Routing of VLSI Chips for Enhanced Reliability," *Proc. of 9th International Symposium on Quality Electronic Design*, pp. 470–475, Mar. 2008.

[5] K. Lu and D. Pan, "Reliability-aware global routing under thermal considerations," *Proc. of 1st Asia Symposium on Quality Electronic Design*, pp. 313–318, July 2009.

[6] K. Sankaranarayanan, S. Velusamy, M. Stan, C. L, and K. Skadron, "A case for thermal-aware floorplanning at the microarchitectural level," *Journal of Instruction Level Parallelism*, vol. 7, 2005.

[7] Y. Han and I. Koren, "Simulated Annealing Based Temperature Aware Floorplanning," *Journal of Low Power Electronics*, vol. 3, 2007.

[8] A. Gupta, N. Dutt, F. Kurdahi, K. Khouri, and M. Abadir, "LEAF: A System Level Leakage-Aware Floorplanner for SoCs," *Proc. of the 2007 Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 274–279, Jan. 2007.

[9] C.-H. Tsai and S.-M. Kang, "Cell-level placement for improving substrate thermal distribution," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, no. 2, pp. 253–266, Feb. 2000.

[10] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. Stan, "HotSpot: a compact thermal modeling methodology for early-stage VLSI design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 5, pp. 501–513, May 2006.

[11] A. Ajami, K. Banerjee, and M. Pedram, "Modeling and analysis of nonuniform substrate temperature effects on global ULSI interconnects," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 6, pp. 849–861, June 2005.

[12] W. C. Elmore, "The Transient Response of Damped Linear Networks with Particular Regard to Wideband Amplifiers," *Journal of Applied Physics*, vol. 19, no. 1, pp. 55–63, Jan. 1948.

[13] J. Black, "Electromigration - a brief survey and some recent results," *Electron Devices, IEEE Transactions on*, vol. 16, no. 4, pp. 338 – 347, apr 1969.

[14] T.-C. Chen and Y.-W. Chang, "Modern floorplanning based on b*-tree and fast simulated annealing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 4, pp. 637 – 650, april 2006.