Compact Thermal Models: Assessment and Pitfalls

Jaeha Kung and Youngsoo Shin
Department of Electrical Engineering, KAIST
Daejeon 305-701, Korea

***Abstract*—Thermal analysis involves solving a differential equation, thus inherently takes a long time. Any thermal optimization techniques require a compact model that can be evaluated fast while accuracy is not sacrificed too much. Several models have been proposed, but their accuracy and runtime have not been fully understood and compared. Three models, namely resistive network, heat diffusion measure, and thermal signature, are studied; a series of experiments reveals that thermal signature is clearly a model of choice. Power density is a key parameter in thermal analysis, and is usually given as a constant value, which corresponds to the average over the region a block occupies and over the time period it operates. The error as a result of approximation is often unnoticed; this is studied in a quantitative way for the first time.**

I. Introduction

Technology scaling has allowed ever increasing integration density along with increasing clock frequency. The increasing power density as a result causes more dissipation of heat. Several effects are known for high temperature: 1) Subthreshold leakage increases exponentially as temperature rises, while temperature itself increases due to larger subthreshold leakage. 2) Electromigration is aggravated [1] and mean-time-to-failure of wire is reduced, e.g. 90% reduction for 27.5°C rise of temperature. 3) Functional errors are observed under nonuniform temperature gradient [2]; difference of temperature across a chip can be as high as about 30°C in high performance microprocessor [3].

It is thus important to estimate temperature accurately, and optimize a design toward lower temperature. Thermal behavior is governed by the well-known heat conduction equation, which originates from the law of conservation of energy:

$$\rho C_p \frac{\partial T(x,y,z,t)}{\partial t} = \nabla[\kappa(x,y,z,t)\nabla T(x,y,z,t)] + g(x,y,z,t), \quad (1)$$

where T is temperature, which is unknown, at particular point (x,y,z) and particular time t, g is power density of a heat source, κ is thermal conductivity, and ρ and C_p are material dependent parameters. A differential equation (1) can be solved by various numerical methods, such as finite difference method (FDM) and finite element method (FEM). But, their application suffers from large runtime, e.g. about 5 hours when FDM is applied to a chip, which is divided into 2.8M imaginary grid cells [4]. There are some faster methods such as the one relying on Green's function [5], but runtime is still large, e.g. several tens of seconds, which is too large to be used in optimization methods such as floorplanning or placement.

Several compact models have been proposed for fast thermal analysis; their accuracy and runtime are assessed in this paper. This is very important for thermal-aware optimization such as floorplanning because only a limited amount of time can be spent in each iteration, e.g. a few ms to determine the location of blocks and compute the total wirelength. Power density g in (1) is usually given as a constant for each functional block for the sake of convenience in computation; it corresponds to average value over the region it occupies and over the time period it operates. This can cause a large error, which surprisingly does not receive enough attention. This is particularly true when a block is very large and its power density is very nonuniform over the area it occupies, and when a power density of a block varies a lot due to aggressive power management being adopted.

The remainder of this paper is organized as follows. In Section II, three compact thermal models are compared and assessed in accuracy and runtime. The effect of taking average value of power density is quantitatively studied in Section III, and we draw conclusion in Section IV.

II. Compact Thermal Models

A. Models

Three compact thermal models are compared in this section: resistive network, heat diffusion measure, and thermal signature.

1) Resistive Network: This model is based on well-known analogy between electrical current and heat flow: heat flow can be described as a current flowing through a thermal resistance, which yields a difference of voltage that is analogous to temperature. Transient analysis requires RC network [6], while resistive network [7] can be used for analysis of steady-state temperature.

The network is derived based on floorplan blocks or imaginary grid cells. In the former approach, several resistances connect blocks in lateral direction and some others connect heat spreader, heat sink, and convective sink-to-air interface in vertical direction. The accuracy of this approach highly depends on the size of blocks [8], i.e. more accuracy is offered for smaller blocks. In the latter, the accuracy can be improved by increasing the number of grid cells, which implies higher runtime. Model order reduction [9] has been proposed to alleviate the situation by reducing the size of resistive network for the same accuracy, but is still not fast enough for practical application.

TABLE I
CORRELATION COEFFICIENT AND RUNTIME OF RESISTIVE NETWORK MODEL

Circuit	Correlation coefficient			
	4×4	8×8	16×16	32×32
nova	0.57	0.85	0.91	0.91
tv80	0.69	0.87	0.93	0.94
wb_dma	0.53	0.76	0.86	0.87
aes	0.47	0.71	0.87	0.89
pci_bridge	0.45	0.64	0.79	0.84
Runtime (ms)	4.41	39.03	185.12	712.74

TABLE II
CORRELATION COEFFICIENT OF HEAT DIFFUSION MEASURE

Circuit	Percentage of hot blocks		
	5%	10%	15%
nova	-0.37	-0.35	-0.63
tv80	-0.12	-0.20	-0.29
wb_dma	-0.22	-0.26	-0.28
aes	-0.19	-0.15	-0.15
pci_bridge	-0.13	-0.23	-0.28

2) Heat Diffusion Measure: Heat diffusion measure [10] tries to estimate the amount of outgoing heat from each block instead of absolute temperature. This uses the observation that the temperature of block is lowered as more amount of heat is diffused out of that block. Heat diffusion is proportional to the difference of power densities of adjacent blocks, i.e. the amount of heat diffusion of block i is given by

$$H_i = \sum_j \left[(d_i - d_j) * \text{shared\_length}\right], \tag{2}$$

where d_i is power density and shared_length is the length of edges of i and j that touch each other. Heat diffusion measure is then given by the sum of H_i over hot blocks:

$$D = \sum_i H_i, \tag{3}$$

Hot blocks are assumed to be those that have higher power densities. Note that if H_i is summed over all blocks, instead of hot blocks, D will be 0. The motivation of heat diffusion measure is that maximum chip temperature will be lowered as more heat is diffused out of hotter blocks, i.e. as D increases. The measure is clearly a crude approximation because only the blocks that touch each other are considered in (2), while adjacent blocks j, even though their edges do not touch that of i, can serve as a heat sink.

3) Thermal Signature: Solving differential equation (1) can be transformed to performing integration by employing Green's function. In steady state, in which temperature does not change over time, (1) is reduced to

$$\nabla^2 T(\mathbf{r}) = -g(\mathbf{r})/\kappa(\mathbf{r}), \tag{4}$$

where $\mathbf{r} = (x,y,z)$ or $\mathbf{r} = (x,y)$ depending on the domain of computation; this is a form of well-known Poisson's equation.

Green's function $G(\cdot)$ is a function that satisfies

$$\nabla^2 G(\mathbf{r}, \mathbf{r_0}) = \delta(\mathbf{r} - \mathbf{r_0}), \tag{5}$$

where δ is the Dirac delta function and $\mathbf{r_0}$ is a point in R^2 or R^3. It can be readily shown that (4) becomes, after some manipulation using (5),

$$T(\mathbf{r}) = -\int_{-\infty}^{\infty} G(\mathbf{r}, \mathbf{r_0}) \frac{g(\mathbf{r_0})}{\kappa(\mathbf{r_0})} d\mathbf{r_0}. \tag{6}$$

Assume that a floorplan is divided into grid cells, and we want to approximate (6) for each cell. Let $\tilde{G}[\,]$ be an approximated Green's function, i.e. its Laplacian is approximately a delta function. We now define a relative estimate of temperature, called thermal signature [11]:

$$TS[i] = \sum_{j \neq i} \tilde{G}[i,j]\, g[j], \tag{7}$$

where $g[j]$ is power density of cell j. Proper choice of $\tilde{G}$ can yield accurate estimate, while evaluation time is kept small.

B. Assessment of Models

Three compact thermal models are experimentally compared in this section. In particular, we assess the accuracy and runtime of each model, in which thermal analysis tool [5] is used as a reference of accuracy.

Five test designs were collected from Opencores [12]. Each design was submitted to logic synthesis tool [13], which also reports the power consumption of each functional block; 100 different floorplans were generated; for each floorplan, maximum temperature is obtained through thermal analysis tool and by using thermal model; correlations are then observed to assess the accuracy of each thermal model.

The correlation coefficient and runtime of resistive network, in which a floorplan is divided into grid cells, are shown in Table I. More accuracy is offered as a floorplan is divided more finely, but at the cost of larger runtime. At least 16×16 is required to ensure correlation coefficient of more than 0.8, which on the other hand needs 185 ms, which is not practical if this model is employed for thermal-aware floorplanning, or any other optimization. Higher runtime is expected for large circuits because more grid cells are involved in the model.

As can be expected, heat diffusion measure is calculated very fast; in less than 100 μs. The accuracy varies depending on how hot blocks are selected. Table II reports the correlation coefficient when hot blocks occupy 5%, 10%, and 15% of total number of blocks. The accuracy is quite unacceptable in all circuits; the correlation coefficient ranges from -0.15 to -0.63; correlations are negative due to the definition of D.

In thermal signature, the proper choice of $\tilde{G}$ is important to achieve both higher accruacy and lower runtime. The following choice is used in the experiment:

$$\tilde{G}[i,j] = \begin{cases} \frac{1}{d_{ij}} & \text{if } d_{ij} \leq R_1 \\ \frac{C}{\sqrt{d_{ij}}} & \text{if } R_1 < d_{ij} \leq R_2 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where d_{ij} is the distance (center to center) between two grid cells i and j; C, R_1, and R_2 are the constants which determine

TABLE III
CORRELATION COEFFICIENT AND RUNTIME OF THERMAL SIGNATURE

Circuit	Correlation coefficient	Runtime (ms)
nova	0.76	0.23
tv80	0.80	0.48
wb_dma	0.85	0.24
aes	0.91	0.86
pci_bridge	0.85	2.50

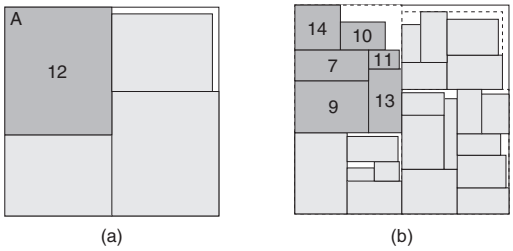

Fig. 1. A chip floorplan viewed in (a) higher hierarchy and (b) lower hierarchy.

the accuracy and the runtime. The reason for the use of two functions ($1/d_{ij}$ and $C/\sqrt{d_{ij}}$) is because the Laplacian of either function alone does not approximate the delta function faithfully (the property of Green's function (5)). By using (8), we calculated maximum thermal signature and compared it to maximum temperature. As shown in Table III, the correlations are quite high; at least larger than 0.76. Along with the high accuracy, computation time to calculate thermal signature is hundreds or thousands of μs. Therefore, it is possible to estimate temperature accurately in thermal-aware optimization process with a little increase in runtime.

III. THE PITFALLS OF THERMAL ANALYSIS

In typical thermal analysis, care needs to be taken both in spatial- and temporal-domain. In spatial domain, an error may arise due to the choice of the size of functional blocks. Power density of a large functional block is effectively the average power density of the blocks contained in it. Any variation of power density across the blocks is thus ignored, which can cause the error. In temporal domain, power density is the average value over the operation period of interest, unless it is given as a function of time during that period. This causes large error as we demonstrate through experiment, in particular because power density varies a lot due to aggressive power management these days.

A. The Errors in Spatial Domain

Fig. 1(a) illustrates a floorplan, which consists of four functional blocks. Let the power density of block A be 12. Fig. 1(b) shows the same floorplan, but the internal blocks that constitute each block of (a) are all depicted, i.e. (b) is a floorplan of blocks that are in lower hierarchy. Now the blocks internal to A have power density from 7 to 14. When thermal analysis is performed in Fig. 1(a), it is assumed that the power density of A is uniform over its area; the result will be clearly different from that obtained by performing thermal analysis on Fig. 1(b), in which power density of A is not uniform.

TABLE IV
DIFFERENCE OF MAXIMUM TEMPERATURE FROM THERMAL ANALYSIS IN DIFFERENT HIERARCHY OF FLOORPLAN

Circuit	Max Temp. (K)		
	Higher	Lower	Diff.
usb_funct	378.5	368.1	10.4
aquarius	362.1	355.2	6.9
aemb	367.0	366.3	0.7
wb_dma	381.1	383.1	-1.9
aes	378.0	367.3	10.7

Five designs were chosen from Opencores [12], which are listed in Table IV; all designs consist of multiple levels of hierarchy. For each design, we picked some hierarchy in a way that the number of functional blocks becomes about 20 to 30; the design was then floorplanned; power consumption of each block was obtained assuming that the transition probability of each circuit input is 0.5, which yields power density to be used by thermal analysis; let us call this a floorplan at higher hierarchy. To obtain a floorplan at lower hierarchy, each block is partitioned into sub-blocks of size about 100 μm^2; sub-blocks are then floorplanned but within the boundary of a parent block; power consumption of each sub-block is also obtained. Thermal analysis [5] is now performed on both floorplans: one at higher hierarchy and another at lower hierarchy. Maximum temperatures are listed in columns 2–3 of Table IV with difference denoted in the last column.

For two examples (aemb and wb_dma), the difference in the last column is small; while it is noticeable in the remainder of examples. This can be understood by looking at the variance of power densities (W/m^3) of sub-blocks within a parent block that consumes most power (thus is likely to be a spot of maximum temperature): 0.86 and 0.85 for aemb and wb_dma, respectively; 5.35, 3.47, and 6.74 for the other three circuits.

B. The Errors in Temporal Domain

Consider a floorplan in Fig. 2(a). The variation of power density over time of each block is also given; the dotted horizontal line indicates an average value, which is used in typical thermal analysis (see g in (4)). Fig. 2(b) shows a thermal map at time 10 ms, when this average power density is used. Thermal analysis when g is given as a function of time can also be done using numerical methods [14]; Fig. 2(c) illustrates the result at the same time of 10 ms. Notice the difference of thermal maps, in particular, in blocks A and B. Assuming average power density implies that A and B, which have higher value than C and D, continuously generate heat. In fact, however, the two blocks rarely generate heat between time 8 ms and 10 ms (because, for example, they are power gated), thus have a chance to be cooled down; this is not reflected in Fig. 2(b) and causes an error.

The same five designs from Section III-A were taken for experiments. For each design, power densities of blocks were obtained [15] in each interval of 2 ms over 10 ms of time

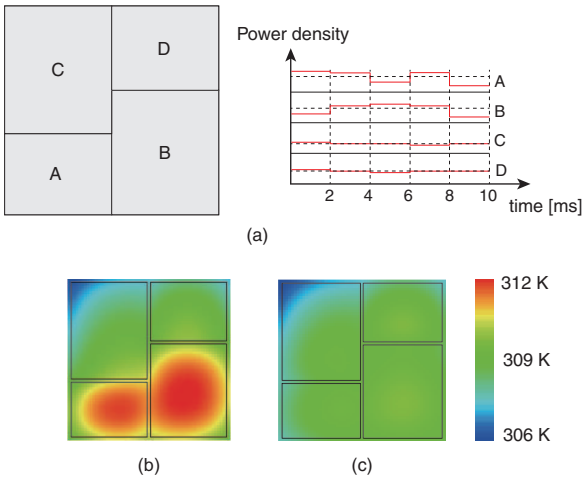

Fig. 2. (a) An example floorplan with variation of power density over time, (b) thermal map at time 10 ms when average power density is used, and (c) thermal map at time 10 ms when variation of power density in (a) is directly used for thermal analysis.

TABLE V
DIFFERENCE OF MAXIMUM TEMPERATURE WHEN AVERAGE POWER DENSITY VALUE AND VARIATION OF POWER DENSITY OVER TIME (DENOTED BY DYNAMIC) ARE USED FOR THERMAL ANALYSIS

Circuit	Max Temp. (K)		
	Average	Dynamic	Diff.
usb_funct	356.5	347.2	9.3
aquarius	348.3	340.5	7.8
aemb	372.1	366.2	5.9
wb_dma	354.0	352.6	1.4
aes	381.4	374.3	7.1

period, similar to Fig. 2(a). Input vectors were arbitrarily adjusted so that some blocks do not receive any value in about 30% of their inputs at each 2 ms interval; this was intended to create a variation of power density. Maximum temperature was then measured from two thermal analyses: one using average power density and another using the variation of power density itself. The results are shown in Table V, with their difference denoted in the last column.

The variation of power density turns out to be small in wb_dma, even though input vectors were adjusted as described before, which explains the small difference. In all other examples, the error when average power density is used deserves attention.

IV. CONCLUSION

Three compact thermal models, resistive network, heat diffusion measure, and thermal signature, were assessed in accuracy and runtime. Resistive network offers high accuracy, but only when large runtime can be tolerated; it thus can not achieve both high accuracy and small runtime. Heat diffusion measure is not useful due to its inaccuracy, even though computation can be done very fast. Thermal siganture turned out to be the model of choice, since it best balances accuracy and runtime.

In typical thermal analysis, power density is given as an average value both in space and time. It was experimentally shown that this causes an error that is worth of paying attention. If a block is very large and its power density is not uniform over the area it occupies, assuming average power density for that block causes an error. Thus, choosing a proper hierarchy during floorplanning becomes important. If power density varies a lot over time due to aggressive power management, it also causes an error.

REFERENCES

[1] 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*, vol. 21, no. 3, pp. 406–411, Sept. 1998.

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

[3] P. Gronowski, W. J. Bowhill, R. P. Preston, M. K. Gowan, and R. L. Allrnon, "High performance microprocessor design," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 5, pp. 676–686, May 1998.

[4] T.-Y. Wang, Y.-M. Lee, and C. Chen, "3D thermal-ADI: an efficient chip-level transient thermal simulator," in *Proc. Int. Symp. on Physical Design*, Apr. 2003, pp. 10–17.

[5] Y. Zhan and S. Sapatnekar, "A high efficiency full-chip thermal simulation algorithm," in *Proc. Int. Conf. on Computer Aided Design*, Nov. 2005, pp. 635–638.

[6] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "HotSpot: a compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. on VLSI Systems*, vol. 14, no. 5, pp. 501–513, May 2006.

[7] 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. 253–266, Feb. 2000.

[8] K. Skadron, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan, "Temperature-aware microarchitecture: modeling and implementation," *ACM Trans. on Architecture and Code Opimization*, vol. 1, no. 1, pp. 94–125, Mar. 2004.

[9] T.-Y. Wang and C. Chen, "SPICE-compatible thermal simulation with lumped circuit modeling for thermal reliability analysis based on modeling order reduction," in *Proc. Int. Symp. on Quality Electronic Design*, Mar. 2004, pp. 357–362.

[10] Y. Han and I. Koren, "Simulated annealing based temperature aware floorplanning," *Journal of Low Power Electronics*, vol. 3, no. 2, pp. 141–155, Aug. 2007.

[11] J. Kung, I. Han, S. S. Sapatnekar, and Y. Shin, "Thermal signature: a simple yet accurate thermal index for floorplan optimization," in *Proc. Design Automation Conf.*, June 2011, pp. 108–113.

[12] "Opencores," http://www.opencores.org/.

[13] Synopsys, "Design Compiler User Guide," Mar. 2007.

[14] E. Choi and Y. Shin, "3-D thermal simulation with dynamic power profiles," in *Proc. Int. Symp. on Circuits and Systems*, May 2008, pp. 2765–2768.

[15] Synopsys, "NanoSim User Guide," Sept. 2008.