Proceedings of the 18th IASTED International Conference
MODELLING AND SIMULATION
May 30-June 1, 2007, Montreal, Quebec, Canada
ISBN Hardcopy: 978-0-88986-663-8/ CD: 978-0-88986-664-5

POWER-DENSITY AWARE FLOORPLANNING FOR REDUCING MAXIMUM ON-CHIP TEMPERATURE

D. Chatterjee and T.W. Manikas
The University of Tulsa
600 South College Avenue, Tulsa, OK-74104, U.S.A.
{debarshi-chatterjee, theodore-manikas}@utulsa.edu

ABSTRACT

With microprocessor power densities escalating rapidly as technology scales below 100nm level, there is an urgent need for developing innovative cooling solutions. In this paper, we introduce the concept of power-density aware thermal floorplanning and demonstrate its efficacy in reducing maximum on-chip temperature. We argue that Compact Thermal Model (CTM) based floorplanners will be hard pressed by time-to-market pressure for placing circuits having large number of modules. To circumvent this problem, we present a novel power-density aware floorplanning technique for reducing the maximum on-chip temperature that has much less runtime compared to CTM based floorplanners. Based on our method we develop a floorplanner that we name COOLER. Instead of using the conventional Simulated Annealing procedure, COOLER uses a highly efficient Multiobjective Evolutionary Algorithm to generate the Pareto-front. Experimental results on MCNC benchmark demonstrate that COOLER is 11x-146x faster than HOTFLOORPLAN. We use HOTSPOT for thermal simulation of the layouts produced by COOLER. Finally, we validate our method by demonstrating that HOTFLOORPLAN solutions lie on the Pareto-front generated by COOLER. It was found that by careful arrangement of components at the architecture level, the average reduction in peak temperature produced by HOTFLOORPLAN and COOLER was 15.1°C and 15.3°C respectively.

KEY WORDS

Computer aided design, genetic algorithm, power-density, thermal, and floorplanning.

1. Introduction

Microprocessor clock frequency doubles each generation and the supply voltage scaling has been unable to compensate the resulting increase in power [1]. If the trend continues, power density will reach 10,000 W/cm^2 by the year 2015 [2]. Increase in power density increases the operating temperature which has severe detrimental effects on the performance of the chip. First, increase in average temperature of a chip deteriorates the chip's reliability due to a phenomenon known as Electromigration (EM). Second, due to the temperature dependence of carrier mobility and interconnect resistivity, the current driving capability of transistors decrease by approximately 4% and the interconnect delay increases by 5% for every 10°C rise in temperature. Furthermore, high temperature increases the risk of a thermal runaway caused by its exponential dependence on leakage power. In order to avoid failures, thermal packages are designed to withstand peak power dissipation. As the peak temperature increases, so does the cost of cooling. The estimated increase in the overall cost of chip due to every Watt of power dissipated above 35-40W is $1/W [3]. Without proper design methodology, this burgeoning cost of cooling will become an impediment for future scaling of devices. The objective of power-density aware floorplanning is to uniformly distribute the power-density, thereby reducing the maximum on-chip temperature and associated packaging cost.

Previous power based methods have failed to reduce maximum on-chip temperature. Skadron et. al. [4] found very small correlation between power and temperature. Temperature-aware design [5] has thus been proposed to address such problems. However, temperature-aware floorplanners using Compact Thermal Model (CTM) suffer from high time-to-market, especially for circuits having large number of modules. Our study shows that there is a high correlation between the steady-state temperature at the center of a block and the distribution of power-density around it. In this paper, we thus introduce the concept of power-density aware thermal floorplanning and study its role in reducing the maximum on-chip temperature.

2. Related Work

One of the pioneering efforts on thermal placement was done by Chu et al. [6]. They developed a matrix synthesis approach for placing square blocks in a way, so as to minimize the maximum sum of powers in all $n \times n$

submatrices. Unfortunately, the floorplanning blocks are not necessarily squares and hence their method cannot be applied to thermal-aware macrocell placement. Tsai et al. [7] proposed a compact substrate thermal model that can be used to estimate temperature profile from a distribution of power dissipating sources on chip. Using their model, it is possible to obtain better thermal distribution for cell level placement without any increase in area. However, for macrocell placement their method gave better temperature distribution at the cost of 30% increase in area. Hung et al. [8] used a Genetic Algorithm (GA) based method for thermal aware floorplanning. Recently, Skadron et al. released HOTFLOORPLAN [9], a temperature aware floorplanner that uses HOTSPOT [10] within a simulated annealing procedure.

However, none of the previous research has directly focused on making power-density uniform. All of the aforementioned methods suffer from high runtime complexity. Moreover, all previous methods reduce the problem of simultaneous optimization of maximum temperature and area to a single objective optimization problem by taking the weighted mean of the two objectives as the cost that is to be minimized. Accordingly, they produce a single optimal solution in a single run. The solution depends on the relative importance of the objectives as defined in the cost function. It requires tedious manipulation of weight terms in the cost function to meet certain specification, such as an upper limit on the area. In this context we make the following contributions: First, we introduce the concept of power-density aware floorplanning. Second, we develop a novel power-density aware technique for macrocell placement that actually reduces the maximum on-chip temperature. Third, we demonstrate that our method is much faster than existing CTM based floorplanners and yet as effective in reducing the maximum on-chip temperature. Fourth, by formulating the thermal placement problem as a Multiobjective Optimization Problem (MOP), we show that a set of Pareto-optimal Solutions can be generated in a single run. Existing methods find solutions that are subsets of the Pareto-optimal Set.

3. Thermal Aware Floorplanning

Given a set of rectangular blocks, thermal-aware floorplanning determines a non-overlapping placement of the blocks to minimize the chip area and maximum on-chip temperature. Floorplan representations have been studied in details. There are two major types of floorplans: slicing floorplans [11] and non-slicing floorplans [12], [13], [14]. Various optimization techniques such as Simulated Annealing (SA) and Genetic Algorithm (GA) have been applied successfully to minimize area. Reducing thermal effects is a relatively nascent area of research. To date, temperature-aware floorplanning has been most effective in reducing peak temperature. We first discuss temperature aware floorplanning technique in brief and then introduce the concept of power-density aware floorplanning.

3.1 Temperature Aware Floorplanning

Temperature-aware floorplanners use an RC circuit to model the steady state and transient temperature at an architectural level. The RC model proposed in [15] has been incorporated in a tool named HOTSPOT [10]. Given the dimensions and relative positioning of all the blocks on chip, HOTSPOT computes the transfer thermal resistance matrix and determines the temperature at the center of each block using:

$$\begin{bmatrix} T_1 \\ T_2 \\ . \\ . \\ T_N \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & . & . & R_{1n} \\ R_{21} & R_{22} & . & . & R_{2n} \\ . & . & . & & . \\ . & . & & . & . \\ R_{m1} & R_{m2} & . & . & R_{mn} \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ . \\ . \\ P_n \end{bmatrix} \tag{1}$$

Where P_i denotes the average power dissipated by the i^{th} block.

3.2 Power-density Aware Floorplanning

The motivation behind power-density aware floorplanning is to model maximum on-chip temperature using a power-density based metric. Since information about the block power-density is readily available, such an approach is computationally efficient. To understand the relation between power-density and temperature, consider the equation for steady-state temperature in a 3D substrate:

$$k\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}\right) + Q(x, y, z) = 0 \tag{2}$$

Subject to the boundary condition:

$$k\frac{\partial T}{\partial n_i} + h_i T = f_i(x, y, z) \tag{3}$$

Where T is the temperature as a function of position; k and r are thermal conductivity (W/m°C) and density (Kg/m^3) of the material respectively, h_i is the heat transfer coefficient of the packaging components (W/m^2°C), Q is the power-density (W/m^3), $\partial / \partial n_i$ represents the differentiation along the outward normal drawn at the boundary surface and f_i is any arbitrary function.

If the power-density Q is uniformly distributed across the spatial coordinates and the initial and boundary conditions are identical for all points, then from symmetry of the differential terms in (2) we can conclude that the temperature distribution will also be uniform. The assumption of uniform initial and boundary condition is justified because any thermal aware methodology, including CTMs must be BICI (Boundary and Initial Condition Independent) [4]. Because power densities are localized and it is not possible to make any changes to the circuitry within a block, it is not possible to obtain an absolutely uniform power-density or temperature distribution simply by redistributing the blocks. However, it is possible to identify high power-density modules as potential candidates for developing hotspots and surround them by whitespace or low power-density module to reduce the effective power-density around potential hotspots.

In order to carry out a mathematical analysis of the problem, let us define the following terms:-

H: Ordered set of high power-density modules (modules having power densities greater than the 80 percentile value) sorted in descending order.

Ω_i : Set of all blocks adjacent to block i $\forall i \in H$

W_{ij}: Shared boundary between $i \in H$ and $j \in \Omega_i$

P_i: Surface power-density of block i

W, L: Width and length of module i respectively.

Let us construct an imaginary rectangle of length $L + 2\Delta x$ and width $W + 2\Delta x$ around block i as shown in Fig. 1.

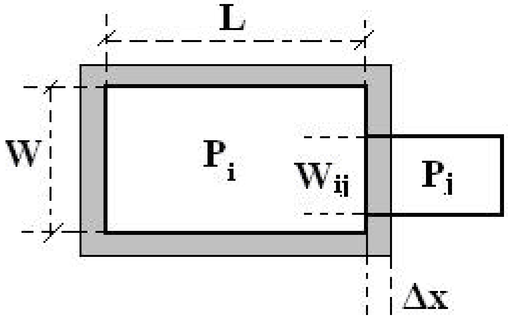

Fig. 1. Diagram showing an imaginary rectangular boundary surrounding block i for computing gradient of power-density reduction.

The effective power-density within the imaginary rectangle is given by:

$$P_{eff} = \frac{P_i WL + \sum_{j\in\Omega_i} W_{ij}\Delta x P_j + O(\Delta x^2)}{(W + 2\Delta x)(L + 2\Delta x)} \qquad (4)$$

The rate at which power-density diminishes with Δx can be calculated from (4) and is given by:

$$\frac{\Delta P}{\Delta x} = \frac{P_i - P_{eff}}{\Delta x} = \frac{1}{WL}\sum_{j\in\Omega_i}(P_i - P_j)W_{ij} \qquad (5)$$

The idea is to determine a layout such that the effective power densities around high power-density modules are minimized. This is equivalent to maximizing the gradient of power-density reduction $\forall i \in H$. In order to reduce the peak temperature, $\Delta P / \Delta x$ for the first element in H, (i.e. the highest power-density block) must be maximized. Once $\Delta P / \Delta x$ for the first element in H has been maximized, we want to select the second element in H and maximize $\Delta P / \Delta x$ for that block. Thus, it will not be correct to develop a scalar thermal objective simply by summing up $\Delta P / \Delta x$ from different blocks in H. This is because, the second highest power-density block might have much greater perimeter, in which case maximizing $\sum \Delta P / \Delta x$ would result in surrounding it with lowest power-density blocks. This will certainly neglect the highest power-density block and hence would fail to reduce the maximum on-chip temperature. To overcome this problem, we define the thermal objective (T.O) to be as follows:

$$T.O = \sum_{i\in H}\left(\frac{1}{S^i}\right)\left(\frac{1}{WL}\right)\sum_{j\in\Omega_i}(P_i - P_j) * W_{ij} \qquad (6)$$

Where S is the scaling factor (~ 10). The entire procedure is summarized in Algorithm 1. From experimental results we find that T.O as defined in (6) has high degree of correlation with maximum and mean temperature across the chip. The correlation coefficients are given in Table1. It is to be noted, that during the calculation of T.O, any whitespace abutting module *i* should be considered as an ordinary block with zero power-density.

Algorithm 1: Procedure for Reducing T_{max}

```
$H \leftarrow$ vector of high power-density blocks in descending order
for all $i \in H$
    find $\Omega_i \leftarrow$ set of blocks adjacent to i , $Scale \leftarrow 1$
    for all $j \in \Omega_i$
        $T.O \leftarrow T.O + (P_{d(i)} - P_{d(j)}) * W_{ij} / Scale$
    end for
    $Scale \leftarrow Scale * S$
end for
```

Table1. Correlation Coefficients between Thermal Objective (T.O) and maximum and mean on-chip temperatures.

BENCHMARKS	T.O VS T_{MAX}	T.O VS T_{MEAN}
apte	0.94	0.67
xerox	0.92	0.91
hp	0.96	0.87
ami33	0.83	0.71
ami49	0.85	0.89

Computing the Area Objective (A.O) is relatively straight-forward. Let P_n be the set of all permutations of $X = \{1,2,...N\}$. We use the sequence pair floorplan representation of Murata et al. [12], where every element in $P_N \times P_N$ represents a valid floorplan. The area objective (A.O) of the floorplan can be defined as follows:

$$A.O = \frac{\sum_{i=1}^{N} A(B_i)}{A_{chip}} \qquad (7)$$

Where $A(B_i)$ is the area of the ith block and A_{chip} is the area of the chip represented by the sequence pair $< X_1, X_2 >$. Chip area can be determined from the chip length and width, which in turn are computed as follows: For every element $x \in X$ it is possible to find a set $M^{bb}(x)$ where:

$$M^{bb}(x) = \{x' \mid x' \text{ is before } x \text{ in both } X_1 \text{ and } X_2\} \qquad (8)$$

Any element $x' \in M^{bb}(x)$ is located left of x in the floorplan. Based on the "left of" constraint imposed by the set, a horizontal constraint graph $G_H(V, E)$ can be constructed. The length of the chip represented by (X_1, X_2) is the distance of the longest path from source to sink in $G_H(V, E)$. The height of the chip can be obtained from the vertical constraint graph $G_V(V, E)$ in a similar manner.

4. Multiobjective Evolutionary Algorithm

Having defined the area and thermal objectives, let us now formulate the problem of thermal-aware floorplanning as a Multiobjective Optimization Problem (MOP). Consider a vector function Φ that maps a decision vector **x** (with decision variables: X_1 and X_2) belonging to the parameter space $P_N \times P_N$ to an objective vector **y** (with objectives: T.O and A.O) belonging to the objective space $\Re^2$. The MOP can then be stated as:

$$\max \ y = \Phi(x)$$

Where $x = < X_1, X_2 > \in P_N \times P_N$

$$y = (TO, AO) \in \Re^2 \quad (9)$$

Where T.O and A.O are given by (6) and (7). We incorporate our power-density aware floorplanning strategy (Algorithm1) in a Multiobjective Evolutionary Algorithm (MOEA) to produce a thermal-aware floorplanner that we name COOLER. To the best of our knowledge, this is the first application of MOEA to thermal floorplanning problem. A variety of MOEA's have been reported in the literature, including VEGA [16], HLGA [17], NSGA [18] and SPEA [19]. Comparative studies based on a large number of test cases have shown that SPEA (Strength Pareto Evolutionary Approach) is best suited for our purpose. Therefore, we use SPEA for solving the MOP. The various steps of SPEA are as follows:

1: Generate a random initial population P and create the empty external nondominated set P'.
2: Compute A.O and T.O for each member in P.
3: Copy nondominated members of P to P'.
4: Remove those members in P' which are covered by other members in P'.
5: If the number of members in P' exceeds a given maximum N', prune P' by means of clustering.
6: Calculate the fitness of each individual in P and P'
7: Select individuals from $P \cup P'$ until the mating pool is filled.
8: Apply crossover and Mutation
9: Stop if the maximum number of generation is reached, else go to step 2.

Operators and Parameters: In our simulation, we used Partially Mapped Crossover (PMX), Swap Mutation and Binary Tournament Selection. We have used crossover and mutation probability values of 0.8 and 0.01 respectively. External and evolving population sizes are set to 40 and 20 respectively.

Clustering and Fitness Assignment Technique:
The clustering technique uses the average linkage method to partition members of P' and then selects the centroid of each cluster as a representative solution. For details on the clustering algorithm, please refer to [19]. The procedure for fitness assignment in SPEA is a two-step process.

Step 1: At first, each member $i \in P'$ is assigned a fitness value (strength) given by:

$$f_i = \frac{n}{N+1} \quad (10)$$

Where n denotes the number of individuals in P that are covered by *i* and N is the size of P.

Step 2: For each member $j \in P$, the sum of the strengths of all nondominated members in P' that covers j has to be computed first. The fitness assigned to j is one greater than the computed sum (to ensure that the members in external population have better fitness values).

$$f_j = 1 + \sum_{i, i f j} f_i \quad (11)$$

5. Simulation Results

We first use HOTFLOORPLAN [9] to place MCNC benchmarks circuits. The results were computed on an Intel Pentium IV laptop running at 1.8 GHz with a 1GB RAM. In the absence of details on power dissipation for MCNC benchmarks, power densities are randomly assigned in the range 1mW/mm^2 to 1W/mm^2. Results in Table 2 show almost cubic growth in runtime as the number of modules increase. It is also observed that reduction in peak temperature for ami33 and ami49 is accompanied by considerable increase in deadspace, which may not fit design specifications. On the other hand for apte and xerox it is possible to bring down the temperature further by relaxing the area constraint. In short, a single run of HOTFLOORPLAN does not give us a clear idea of the trade off surface or the Pareto-optimal solutions at our disposal.

Next we used COOLER to generate thermal-aware layouts for all MCNC benchmarks. One such layout for apte is shown in Fig 2(a). We found that the runtime for placement varies from 23 sec for apte to 401 sec for ami49. Thus our method is around 11-146 times faster than HOTFLOORPLAN. Fig. 3 shows the comparison of runtimes for COOLER and HOTFLOORPLAN. This clearly demonstrates the superiority of our method over CTM based floorplanners especially when the number of modules is large.

We use HOTSPOT [10] to analyze the spatial distribution of temperature (at the block level granularity) for the layouts produced by COOLER. The temperature distribution for the apte layout in Fig. 2(a) is plotted in Fig 2(b) using MATLAB. For each MCNC benchmark, COOLER produces a Pareto-optimal solution set. The maximum and mean temperature for each layout in the

Table 2. Percentage Deadspace, Maximum and Mean on-chip temperatures and runtime for HOTFLOORPLAN

MCNC Benchmarks	Percentage Deadspace	T_{max} (K)	T_{mean}(K)	Runtime (sec)
apte	3.27	348.90	343.91	257
xerox	7.83	337.5	329.01	399
hp	13.14	349.6	348.01	501
ami33	44.89	327.5	327.02	13526
ami49	17.43	341.8	336.21	58838

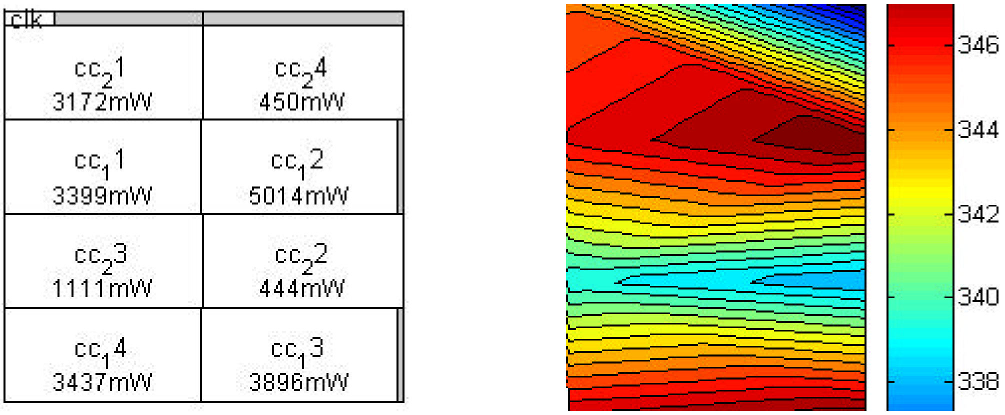

Fig. 2. (a) Thermal Aware Layout of apte generated by COOLER (b) Spatial Distribution of temperature for the same layout.

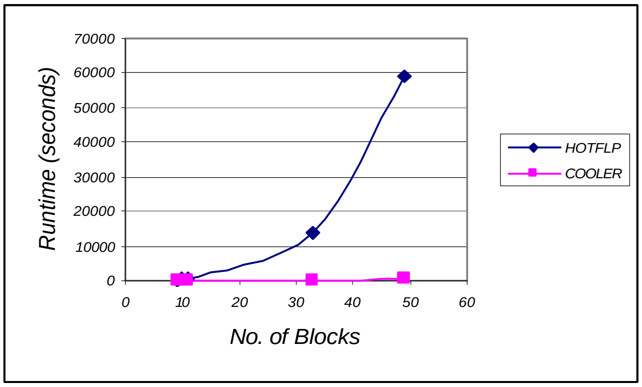

Fig. 3. Runtime versus the number of modules for COOLER and HOTFLOORPLAN.

Pareto-optimal set is extracted using HOTSPOT and plotted in Fig. 4. This plot delineates the Pareto-front for each benchmark by joining the Pareto-optimal solutions using a smooth curve. The maximum temperature and percentage deadspace corresponding to the layouts generated by HOTFLOORPLAN are also plotted on the same figure. It is observed that the solutions produced by HOTFLOORPLAN lie on the Pareto-front generated by COOLER. We thus conclude that COOLER is as effective in reducing T_{max} as CTM based temperature-aware floorplanners and yet has very less runtime. Moreover, our method allows the designer to explore various design possibilities without having to modify the objective function in the source code. For example, Fig. 4 indicates that T_{max} in apte can be reduced from 357K to 347K by increasing the deadspace by 7%. In short, by using an MOEA as opposed to conventional Simulated Annealing or single objective GA it is possible to get a complete picture of the trade-off surface. Fig. 5 shows that COOLER and HOTFLOORPLAN are capable of reducing T_{max} by 15.3°C and 15.1°C on an average.

6. Conclusion

Power-density aware thermal floorplanning has immense potential to reduce peak temperature and thereby enhance performance and reliability of current and future generation ICs. Modeling the maximum on-chip temperature using a power density based metric enables reducing the runtime considerably without sacrificing the quality of the solution. The Multiobjective framework helps to study the trade-offs between the various objectives. The technique introduced for reducing the peak temperature has very low time-to-market and hence will find potential application in industrial IC design.

7. Acknowledgement

The authors would like to thanks Karthik Sankaranarayanan for his help in troubleshooting HOTSPOT. We also wish to thank Dr. D. R. Doty and Partha Mukherjee for their helpful comments and suggestions.

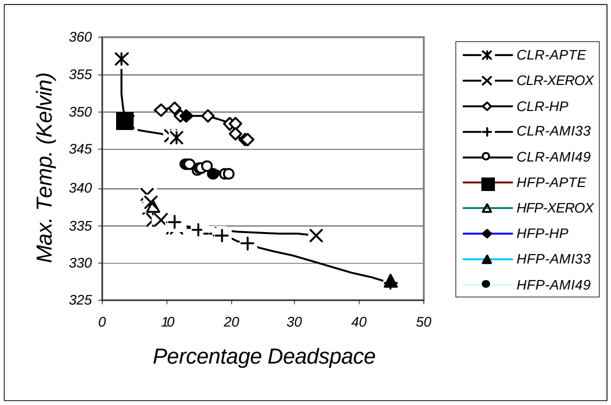

Fig. 4. Pareto-optimal solutions generated by COOLER (CLR) and solutions produced by HOTFLOORPLAN (HFP) for MCNC benchmarks.

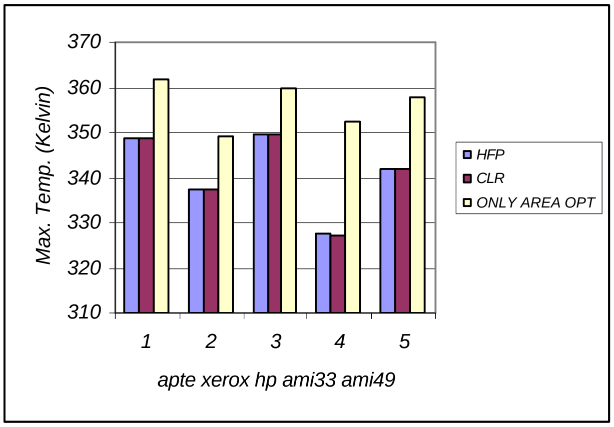

Fig. 5. Reduction in Tmax by COOLER (CLR) and HOTFLOORPLAN (HFP) as compared to non-thermal floorplanner.

References

[1] SIA. International Technology Roadmap for Semiconductors, 2001.

[2] http://www.intel.com/technology/magazine/computin g /platform-2015-0305.htm

[3] Borkar, S. Design Challenges of Technology Scaling. *IEEE Micro*, Jul.-Aug. 1999, 23-29.

[4] Skadron, K., Stan, M.R., Huang, W., Velusamy, S., Sankaranarayanan, K., Tarjan, D. Temperature-Aware Microarchitecture. *30th Intl. Symp. on Comp. Architecture*, June 2003, 2-13.

[5] Huang, W., Stan, M.R., Skadron, K., Sankaranarayanan, K., Ghosh, S., and Velusamy, S. Compact Thermal Modeling for Temperature-Aware Design, *DAC*, June 2004.

[6] Chu, C. N., and Wong, D. F. A Matrix Synthesis Approach to Thermal Placement. *IEEE Trans. on CAD of IC and Systems*, *17*, 11, Nov. 1998, 1166-1174.

[7] Tsai, C.-H., and Kang, S.-M. Cell-Level Placement for Improving Substrate Thermal Distribution. *IEEE Trans. on CAD of IC and Systems*, *19*, 2, Feb. 2000, 253-266.

[8] Hung, W.-L., Xie, Y., Vijaykrishnan, N., Addo-Quaye, C., Theocharides, T., and Irwin, M. J. Thermal Aware Floorplanning using Genetic Algorithms. *6th ISQED*, 2005, 634-639.

[9] K. Sankaranarayanan, S. Velusamy, M.R. Stan, and K. Skadron. A Case for Thermal-Aware Floorplanning at the Microarchitectural Level. *Journal of Instruction-Level Parallelism*, Sept. 2005.

[10] http://lava.cs.virginia.edu/HotSpot/index.htm

[11] D.F. Wong & C.L Liu, A new Algorithm for Floorplan Design, *Design Automation Conference*, 1986, 101-107.

[12] H. Murata, K. Fujiyoshi, S. Nakatake and Y. Kajitani, VLSI Module Placement Based on Rectangle-Packing by Sequence Pair, *IEEE Transactions on Computer Aided Design, Vol. 15*(12), 1996, 1518-1524.

[13] P.-N Guo, C. -K. Cheng and T. Yoshimura, An O-tree Representation of Non-slicing Floorplan and its Applications, *Design Automation Conference*, 1999, 268-273.

[14] Y.-C. Chang, Y.-W. Chang, G.-M. Wu and S.-W. Wu, B*-Trees: A new Representation for Non-Slicing Floorplans, *DAC*, 2000, 458-463.

[15] Skadron, K., Abdelzaher, T., and Stan, M. R. Control-Theoretic Techniques and Thermal-RC Modeling for Accurate and Localized Dynamic Thermal Management. *8th Intl. Symp. on High Performance Comp. Architecture*, Feb. 2002, 17-28.

[16] Schaffer, J. D. Multiple Objective Optimization with Vector Evaluated Genetic Algorithms. *Genetic Algorithms and their Applications: First International Conference on Genetic Algorithms*, Lawrence Erlbaum, Mahwah, NJ, 1985, 93-100.

[17] Hajela, P., and Lin, C.-Y. Genetic Search Strategies in multicriterion optimal design. *Structural Optimization*, *4*, New York: Springer, June 1992, 99-107.

[18] Srinivas, N., and Deb, K. Multiobjective Optimization using Nondominated Sorting in Genetic Algorithms. *Evolutionary Computation*, *2*, 3, 1994, 221-248.

[19] Zitzler, E., and Thiele, L. Multiobjective Evolutionary Algorithms: A Comparative Case Study and Strength Pareto Approach. *IEEE Trans. on Evolutionary Computation*, *3*, 4, Nov. 1999, 257-271.