Efficient Thermal via Planning Approach and Its Application in 3-D Floorplanning

Zhuoyuan Li, Xianlong Hong, *Fellow, IEEE*, Qiang Zhou, Shan Zeng, Jinian Bian, Wenjian Yu, *Member, IEEE*, Hannah H. Yang, Vijay Pitchumani, and Chung-Kuan Cheng, *Fellow, IEEE*

***Abstract*—In this paper, we investigate thermal via (T-via) planning during three-dimensional (3-D) floorplanning. First, we consider the temperature constrained T-via planning (TVP) problem on a given 3-D floorplan. Second, we integrate dynamic TVP into 3-D floorplanning process. Our main contribution and results can be summarized as follows. We solve the temperature constrained TVP problem by solving a sequence of simplified interlayer and intralayer TVP subproblems. Each subproblem is formulated as convex programming problem and we derive nearly optimal solution for detailed T-via distribution. Based on the TVP solution, we implement the integrated TVP and 3-D floorplanning algorithm in a two-stage approach. Before floorplanning, blocks are assigned into different layers by solving a sequence of knapsack problems. During floorplanning, T-vias are allocated with white space redistribution to optimize T-via insertion. Experimental results show that our TVP approach can reduce T-vias by 12% compared with a recent published work (J. Cong and Y. Zhang, "Thermal via planning for 3-D ICs," in *Proc. Int. Conf. Comput.-Aided Des.*, Nov. 2005, pp.745–752). Compared with the postfloorplanning optimization approach, integrating TVP into floorplanning process can reduce T-vias by 16% with 21% runtime overhead.**

***Index Terms*—Floorplanning, optimization, thermal, very large scale integration (VLSI).**

I. Introduction

With the exponential growth of circuit complexities, the system complexity continues to increase and physical design is getting more and more difficult. Cycle time optimization has become one of the most important issues in the design of highly integrated circuits. As three-dimensional (3-D) technology can promise higher integration density, lower interconnection complexity, and delay, it is being viewed as a potential alternative that cannot only alleviate interconnect delay problem, increase transistor packing density and reduce chip area significantly, but also inspire a new generation of circuit design concepts [1], [2]. Despite its advantages over traditional two-dimensional (2-D) ICs, the heat dissipation has become an extremely important issue in 3-D IC design [3]. The thermal effects are expected to be exacerbated by the reduction in chip size, resulting in a sharp increase in the power density. Moreover, the heat is typically conducted through the silicon substrate to the package, and then to the ambient by a heat sink. With multiple device layer designs, devices in the upper stack will cause a significant fraction of the heat, which can cause great degradation in device performance and reduction in chip reliability.

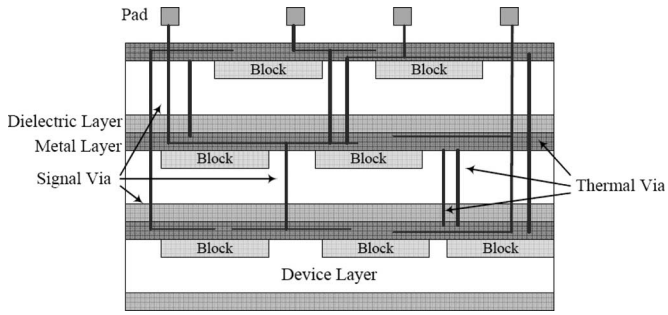


Fig. 1. T-via in 3-D IC stack.

There is scattered literature on thermal optimization during different physical design stages including 3-D floorplanning [4], [5], 3-D placement [8]–[11], [31], and routing [12], [14], [15]. However, even with a complicated thermal-oriented algorithm to improve heat resource distribution in 3-D integration, the maximal temperature is still much higher than that of 2-D design due to low thermal conductivity between different device layers [5], [9]. Therefore, advancement in cooling and packaging technologies is necessary to maintain acceptable chip temperatures. One effective way of reducing circuit temperature is to incorporate "dummy thermal vias" into 3-D ICs to mitigate thermal issues by lowering the thermal resistance between different device layers [16], [17]. These thermal vias are drilled through device layers as additional electrically insulated vias besides the through-the-silicon signal vias as shown in Fig. 1. Thermal vias exist only for temperature reduction and have no connection to metal signal wires.

Under current technology, thermal vias (T-vias) are costly to fabricate. On the other hand, through-the-silicon T-via pitch is usually much larger than that of regular metal wires. They are generally obstacles for routing and large amount of T-vias would lead to serious congestion problem in 3-D ICs. Therefore, T-via planning (TVP) algorithms are needed to minimize the total number of inserted T-vias while placing them in hot areas to make the greatest impact. Previous work [13], [14] formulate and solve the TVP problem as a postplacement

Manuscript received February 13, 2006; revised May 16, 2006 and July 17, 2006. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Award 90407005 and in part by Intel Corporation. This paper was recommended by Guest Editor J. Hu.

Z. Li, X. Hong, Q. Zhou, S. Zeng, J. Bian, and W. Yu are with the Computer Science and Technology Department, Tsinghua University, Beijing 100084, China (e-mail: we97@mails.tsinghua.edu.cn; hxl-dcs@tsinghua.edu.cn).

H. H. Yang is with the Strategic Computer-Aided Design (CAD) Laboratory, Intel Corporation, Hillsboro, OR 97124 USA.

V. Pitchumani is with Intel Corporation, Santa Clara, CA 95054 USA.

C.-K. Cheng is with the Computer Science and Engineering Department, University of California, San Diego, CA 92093 USA.

Digital Object Identifier 10.1109/TCAD.2006.885831

optimization stage for further improving the thermal distribution based on thermal-driven floorplanning/placement results. In [13], T-vias are modeled to be arranged in specific T-via regions and used to adjust their effective thermal conductivities. The T-via distribution is determined based on explicit thermal profiling and the T-via placement method makes iterative adjustments to T-via distribution in order to achieve a desired maximum temperature objective.

In [14], the temperature constrained TVP problem is formulated as a constrained nonlinear programming problem (NLP) based on the resistive thermal model. An efficient heuristic method multilevel alternating direction via planning (m-ADVP) is developed to solve the constrained NLP problem by solving a sequence of simplified interlayer and intralayer TVP subproblems. The interlayer TVP is formulated as a constrained convex programming problem (CP) and the authors derive an analytical solution for the constrained CP problem. The intralayer TVP is based on heat propagation technique and their method also makes iterative adjustments to T-via distribution. It is shown that their algorithm is much faster than the direct solution to the constrained NLP problem. Moreover, the maximal temperature can be brought down to desired threshold by iterative T-via distribution adjustment and heat flow analysis process.

These TVP results may be greatly affected by original floorplanning and placement results. For example, T-vias should be inserted at the white spaces between macroblocks so that heat generated by these blocks could flow to the neighboring white spaces and through T-vias to the bottom layer. Generally, the hot areas are occupied by compactly placed blocks so it is difficult to get enough white space at desired regions that T-vias should be inserted. The maximal temperature may not be brought down to desired threshold with T-via insertion on initial floorplanning result. The floorplan may need to be modified to fit T-via insertion, which would cause degradation on overall packing area and total wirelength. Therefore, integrating TVP with thermal driven floorplanning algorithms are beneficial for both T-via insertion and floorplanning solution quality.

However, previous TVP algorithms determine T-via distribution based on explicitly thermal profiling result [13] or complicated heat propagation analysis iterations [14]. Integrating these algorithms into 3-D floorplanning process is too time consuming. On the other hand, experimental results show that the simulated annealing (SA) process for thermal-oriented 3-D floorplanning is not stable enough, especially when we take the T-via effect into consideration. Integrating these complicated TVP approaches into the SA process would seriously affect the convergence speed.

In this paper, we investigate TVP phase during 3-D floorplanning. First, we derive a nearly optimal analytical solution for detailed T-via distribution on a given 3-D foorplanning result. Similar to [14], the temperature-constrained TVP problem is transformed into a sequence of simplified interlayer and intralayer TVP subproblems. 1) The interlayer TVP subproblem is formulated as a CP and we derive the solution for ideal interlayer T-via distribution. 2) With the heat flow analysis technique, the intralayer TVP subproblem is formulated as a NLP with maximal thermal gradient constraints. Then, it is transformed into a simplified CP and we derive a nearly optimal solution for this problem. With our method, detailed T-via distribution can be determined without explicit temperature profiling.

Based on the analytical solution, we integrate dynamic TVP into thermal-oriented 3-D floorplanning process. The integrated 3-D floorplanning and TVP (3-DFP-TVP) algorithm is implemented in a two-stage approach. Before floorplanning, blocks are assigned to different device layers. Then, floorplans of all these layers are generated in a SA process. The new two-stage 3-DFP-TVP approach scales down the much enlarged solution space due to multiple device layer structure. It not only results in a much faster convergence speed and lower design complexity, but also interacts well with our TVP approach. 1) Based on the analytical solution for interlayer T-via distribution, the interlayer partition of blocks to minimize total number of T-vias can be determined by solving a sequence of simplified knapsack problems. 2) During floorplanning, detailed T-via distribution is determined with above intralayer TVP method. White space redistribution method is applied to further improve T-via insertion without sacrifice on packing area.

The rest of this paper is organized as follows: Section II describes the problem formulations for the TVP and 3-D floorplanning problems. Section III introduces the resistive thermal model and effect of T-vias. The temperature-constrained TVP problem is formulated and solved in Section IV. Section V presents the integrated TVP and 3-D floorplanning algorithm. Experimental results are reported and compared in Section VI. Section VII is the conclusion.

II. Problem Formulation

Given a set of rectangular blocks $B = \{B_1, B_2, \ldots, B_n\}$, each rectangular block B_i is defined by a tuple (h_i, w_i), where h_i and w_i are the height and the width of B_i, respectively. Assume the total number of stacked layer is k. Let (x_i, y_i, l_i) denote the coordinate of the left lower corner of the rectangle B_i, where $1 \leq l_i \leq k$, $l_i \in N$. A 3-D floorplan F is an assignment of (x_i, y_i, l_i) for each B_i such that no two blocks overlap, while white spaces are reserved between blocks for interlayer interconnects and T-via insertion. T-vias are arranged in the white space between blocks.

In this paper, two different schemes for TVP problem are investigated. In the first scheme, we formulate and solve the temperature-constrained TVP problem. Given a 3-D floorplan and a maximal temperature threshold, we determine desired T-via distribution to satisfy the maximal temperature constraint. For reasons related to fabrication cost and limited routing resources, the total number of inserted T-vias are to be minimized. The TVP problem is addressed in Section IV.

In the second scheme, we integrate TVP with 3-D floorplanning. T-via distribution is dynamically updated during floorplanning process. The goal of a thermal-oriented 3-D floorplanning with TVP algorithm is to minimize the total number of T-vias, chip area, and total wirelength with the constraint of a given maximal temperature. This problem is addressed in Section V.

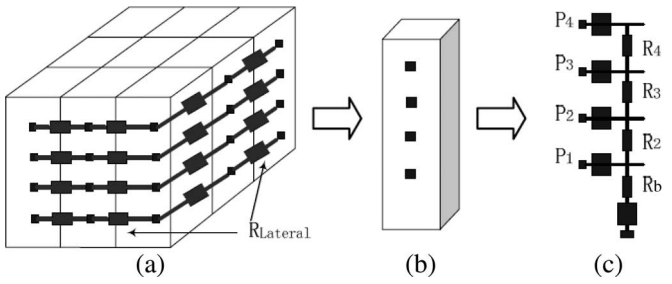


Fig. 2. Compact resistive thermal model. (a) Tile stack array. (b) Single tile stack. (c) Resistance.

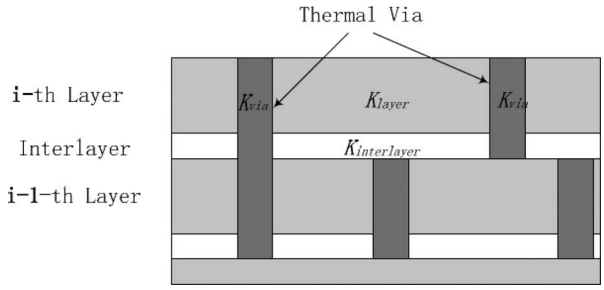


Fig. 3. T-via density and thermal conductivity.

III. T-VIA MODELING

To determine T-via distribution for thermal optimization, thermal analysis is needed to calculate temperature distribution. In the following sections, we will introduce the resistive thermal model and T-via impact model.

A. Resistive Thermal Model

Generally, on-chip temperature distribution can be calculated by numerical thermal model such as finite-difference method (FDM) [22], [23] and finite-element analysis [8]. Due to the large problem size in very-large-scale-integration systems, some modification and simplification, such as multigrids computation [24] and analytical Green function method [25] are proposed to avoid heavy computational load of such precise models.

In this paper, we use a recently proposed resistive thermal model for thermal profiling [20], which can explicitly model the impact of T-vias. The 3-D circuit stack is divided by a 2-D array of tile stacks as shown in Fig. 2(a). Each tile stack is composed of several vertically stacked tiles, one from each device layer as shown in Fig. 2(b). These tile stacks are connected by normalized lateral thermal resistances R_{Lateral}. Within each tile stack, a thermal resistor R_i is modeled for the ith device layer, while thermal resistance of the bottom layer and silicon substrate is modeled as R_{b} as shown in Fig. 2(c). R_{Lateral} and R_{b} are fixed resistances decided by the technology parameters and the size of tiles. R_i is variable resistance that will be affected by the density of T-vias inside that tile.

For boundary conditions, the four sides and top of the 3-D chip are treated as adiabatic, since the chip is usually packaged in thermal insulated materials. The bottom side is isothermal of constant room temperature. In the resistive thermal model, the values of thermal resistances are calculated by accurate finite-element-analysis (FEA)-based simulation. This was accomplished by solving the resistive network utilizing known temperature values obtained from thermal simulations conducted and minimizing the error squared values [21]. When applying this model for TVP during 3-D floorplanning, thermal resistance values can be estimated before routing since there are only a few block-level global nets and routing these nets have a little impact on the thermal distribution. In our implementation, we assume that these nets are distributed uniformly in the chip area and calculate the thermal resistances.

The temperature at each node of the network is analogous to the voltage on the node and the power density value at each node is treated as a current source. Given the positions of macroblocks and T-vias, the linear system of thermal resistive network can be solved by a linear solver.

B. T-via Modeling

In this paper, we use the same T-via modeling as [14]. As shown in Fig. 3, T-via can go through multiple device layers or go between adjacent layers if it is not overlapped with blocks. The T-via density in the ith layer is given by the following equation:

$$m_i = n_i A_{\text{via}}/S$$

where n_i is the total number of T-vias in the ith layer. A_{via} is the cross-sectional area of each T-via. S is the packing area of stacked layers.

The relation between T-via density and the effective vertical thermal conductivity of each layer and interlayer is given by

$$\begin{aligned} K_{i1} &= m_i K_{\text{via}} + (1 - m_i) K_{\text{layer}} \\ K_{i2} &= m_i K_{\text{via}} + (1 - m_i) K_{\text{interlayer}} \end{aligned} \tag{1}$$

where K_{via} is the thermal conductivity of the via material, K_{layer} is the thermal conductivity of each layer without any T-vias, and $K_{\text{interlayer}}$ is the thermal conductivity of each interlayer without any T-vias as shown in Fig. 3.

With the duality between heat transfer and electrical current flow, any thermal resistance can be regarded as an electrical conductor. The relations between thermal resistance R and thermal conductivity K can be expressed as [3]

$$R = \frac{\alpha}{K} \times \frac{l}{s} \tag{2}$$

where l and s are the length and cross-sectional area of the thermal conductor. α is a scaling factor and it is inversely proportional to the density of heat resources. If the whole chip area is occupied by heat resources $\alpha = 1$.

Using (2), the thermal resistance between the ith layer and the $i-1$th layer is modeled as the series thermal resistances of the ith layer and interlayer

$$R_i = \frac{\alpha}{K_{i1}} \times \frac{L_{\text{layer}}}{S} + \frac{\alpha}{K_{i2}} \times \frac{L_{\text{interlayer}}}{S} \tag{3}$$

where L_{layer} and $L_{\text{interlayer}}$ are the thickness of each layer and interlayer, respectively. It is shown that R_i is inversely proportional to the T-via density inside each layer.

TABLE I
SIMULATION RESULTS. THE LAYER THICKNESS IS SET TO BE 20 μm. THE LAYER THICKNESS IS SET TO BE 2 μm. THE THERMAL CONDUCTIVITIES OF LAYER AND INTERLAYER ARE SET TO BE 119 AND 1.1 W/mC, RESPECTIVELY. THE FOUR SIDES AND TOP OF THE TILE STACKS ARE TREATED AS ADIABATIC

Tile Stack (um)	α	Simulation Resistance	Predicted Resistance	Percent Error
50×80	1.25	611	620	-2.5%
50×100		499	496	+0.1%
100×100		257	248	+3.6%
100×130		195	191	+2.1%
200×200		66	62	+6.4%

We use a detailed FDM simulator [22] to validate the analytical model for thermal resistance calculation. Heat resources are placed on the ith layer randomly and they occupy 80% chip area. We calculate the temperature rise on the ith layer and the $i-1$th layer with the FDM simulator. Based on the simulation results, the effective thermal resistance of adjacent layer should be

$$R_i = \frac{T_i - T_{i-1}}{P_i}$$

where T_i is the temperature on the ith layer. P_i is the total power consumption on the ith layer. Table I shows the results compared with predicted thermal resistances calculated with (3). It is shown that the predicted resistance by the analytical model is indistinguishable from the FDM simulation results.

IV. T-VIA PLANNING (TVP)

Given a 3-D floorplan and a maximal temperature threshold, T_0, the temperature constrained TVP problem could be expressed as

$$\min \sum_{i=2}^{k} m_i$$
$$\text{s.t.} \quad \max(T_{1,\max}, T_{2,\max}, \ldots, T_{k,\max}) \le T_0$$

where $T_{i,\max}$ is the maximal temperature on the ith layer. As shown in [14], this problem is a constrained NLP problem. In order to efficiently solve the T-via number minimization problem, the authors propose a two-step relaxation for the original problem: Interlayer TVP distributes T-vias to different layers and Intralayer TVP distributes T-vias inside each layer. Experimental results show that their algorithm is much faster than the direct solution to the NLP formulation for via planning with very similar solution quality. Below, we apply this two-step relaxation to the TVP problem and derive analytical solutions for these subproblems.

A. Interlayer TVP

To determine interlayer T-via distribution, we assume that the temperature distribution and T-via distribution inside each device layer are uniform. With the analytical temperature model in [3], the temperature of the ith layer can be calculated as

$$T_i = R_{\mathrm{b}} \sum_{j=1}^{k} P_j + \sum_{l=2}^{i} \left(R_l \sum_{j=l}^{k} P_j \right) + T_{\mathrm{amb}} \tag{4}$$

where T_{amb} is the ambient temperature. This equation indicates that the top layer would have the maximal temperature. With (1), (3), and (4), the temperature constrained vertical TVP problem is formulated as

$$\begin{aligned}
&\min \sum_{i=2}^{k} m_i \\
&\text{s.t.} \ \ (\text{a}) \ 0 \le m_i \le C_i \\
&\qquad (\text{b}) \ \max(T_1, \ldots, T_k) = T_k \le T_0 \\
&T_k = \sum_{i=2}^{k} \left(\left(\frac{\theta_{\mathrm{layer}}}{1+\lambda_{\mathrm{layer}} m_i} + \frac{\theta_{\mathrm{interlayer}}}{1+\lambda_{\mathrm{interlayer}} m_i} \right) \sum_{j=i}^{k} P_j \right) \\
&\qquad + R_{\mathrm{b}} \sum_{j=1}^{k} P_j + T_{\mathrm{amb}}, \\
&\lambda_{\mathrm{layer}} = \frac{K_{\mathrm{via}} - K_{\mathrm{layer}}}{K_{\mathrm{layer}}}; \quad \theta_{\mathrm{layer}} = \frac{\alpha L_{\mathrm{layer}}}{K_{\mathrm{layer}} S}; \\
&\lambda_{\mathrm{interlayer}} = \frac{K_{\mathrm{via}} - K_{\mathrm{interlayer}}}{K_{\mathrm{interlayer}}}; \\
&\theta_{\mathrm{interlayer}} = \frac{\alpha L_{\mathrm{interlayer}}}{K_{\mathrm{interlayer}} S}
\end{aligned}$$

where C_i is the T-via capacity of the ith layer. Constraints (a) specify the maximal amount of T-vias assigned to each layer. As T-vias are to be inserted to white space in a floorplan, there must be an upper bound for T-via capacity in practical IC design. In our algorithm, we ignore the range constraint to make the problem easier to solve. Experimental results show that resulting m_i is within the range since in most case the range is fairly loose [14].

However, it is inefficient to solve this NLP directly when considering constraint (b). Note that the coefficients of the variables have symmetric forms. To accelerate the computation, we treat the layer and the interlayer as a unity. It has an average thermal conductivity K_{avg} and its thickness is the sum of the thicknesses of layer and interlayer. The value of K_{avg} is related to the thickness and thermal conductivity of layer and interlayer materials. With the simplified model, the temperature constrained vertical TVP problem can be rewritten as

$$\begin{aligned}
&\min \sum_{i=2}^{k} m_i \\
&\text{s.t.} \quad \sum_{i=2}^{k} \left(\frac{\theta}{1+\lambda m_i} \sum_{j=i}^{k} P_j \right) + R_{\mathrm{b}} \sum_{j=1}^{k} P_j + T_{\mathrm{amb}} \le T_0, \\
&\lambda = \frac{K_{\mathrm{via}} - K_{\mathrm{avg}}}{K_{\mathrm{avg}}}; \qquad \theta = \frac{\alpha (L_{\mathrm{layer}} + L_{\mathrm{interlayer}})}{K_{\mathrm{avg}} S}.
\end{aligned} \tag{5}$$

The LHS of above constraint is a function of T-via density inside each layer. Let

$$F(m_2,\ldots,m_k)=\sum_{i=2}^{k}\left(\frac{\theta}{1+\lambda m_i}\sum_{j=i}^{k}P_j\right).$$

Lemma 1: Function F is convex on the interval [0, 1].

Proof: The partial derivative of F is

$$\frac{\partial F}{\partial m_i}=-\sum_{j=i}^{k}P_j\times\frac{\theta\lambda}{(1+\lambda m_i)^2}.$$

The second derivative of F is

$$\nabla^2F(m_2,\ldots,m_k)=\begin{bmatrix}\frac{\lambda^2\theta}{(1+\lambda m_2)^3}\sum_{j=2}^{k}P_j & \cdots & 0\\ \vdots & \cdots & \vdots\\ 0 & \cdots & \frac{\lambda^2\theta}{(1+\lambda m_k)^3}\sum_{j=k}^{k}P_j\end{bmatrix}.$$

It is shown that ∇^2F is positive definite, so F is a convex function.

This NLP in (5) is a CP. It can be solved directly with the Karush–Kuhn–Tucker (KKT) optimal condition. Let

$$f=\sum_{i=2}^{k}m_i+\mu\left(\Delta T-\sum_{i=2}^{k}\left(\frac{\theta}{1+\lambda m_i}\sum_{j=i}^{k}P_j\right)\right)$$

where $\Delta T=T_0-T_{\mathrm{amb}}-R_{\mathrm{b}}\sum_{j=1}^{k}P_j.$

The KKT optimal condition requires

$$\partial f/\partial m_i=0,\qquad 2\le i\le k.$$

By solving the above equations, we derive the analytical solution for this problem

$$m_2=\left(\frac{\theta\sqrt{Q_2}\sum_{i=2}^{k}\sqrt{Q_i}}{T_0-T_{\mathrm{amb}}-R_{\mathrm{b}}Q_1}-1\right)\Big/\lambda$$

$$1+\lambda m_i:1+\lambda m_{i-1}=\sqrt{Q_i}:\sqrt{Q_{i-1}},\qquad 3\le i\le k$$

where
$$Q_i=\sum_{l=i}^{k}P_l. \tag{6}$$

These equations determine the optimal interlayer T-via distribution only if the temperature distribution inside each layer is uniform, which results in uniform intralayer T-via distribution. It is too difficult to generate such a temperature distribution so this solution could not be applied to determine the detailed T-via distribution. However, (6) determines the desired T-via densities on each layer to satisfy the maximal temperature constraint. After determining the T-via densities on each layer, the desired maximal temperature on the ith layer with T-via insertion, $T_{i,\max}$, could be calculated by (4). The detailed T-via distributions inside each layer should result in temperature distributions to satisfy this constraint. In the following section, we address on this problem and propose an efficient approach to determine intralayer via distribution.

B. Intralayer TVP

As T-vias are used for better thermal conductivity between different device layers, optimizing the maximal on-chip temperature could be replaced by minimizing the maximal thermal gradient between adjacent device layers. The desired maximal vertical thermal gradient between the ith layer and the $i-1$th layer could be expressed as

$$\Delta T_i=T_{i,\max}-T_{i-1,\max}.$$

The vertical thermal gradients of tiles on the ith layer should be upper bounded by the maximal thermal gradient. The vertical thermal gradient in tile t_{ik} can be expressed as

$$\Delta T_{ik}=I_{ik}R_{ik} \tag{7}$$

where I_{ik} is the vertical heat flow from the ith layer to the $i-1$th layer inside t_{ik}. R_{ik} is the vertical thermal resistance inside t_{ik}. With the T-via modeling introduced in Section III-B, the relationship between T-via density inside t_{ik} and R_{ik} is given by

$$R_{ik}=\frac{\alpha(L_{\mathrm{layer}}+L_{\mathrm{interlayer}})/A_{\mathrm{t}}}{m_{ik}K_{\mathrm{via}}+(1-m_{ik})K_{\mathrm{avg}}} \tag{8}$$

where A_{t} is the area of each tile. The vertical heat flow I_{ik} can be calculated by the resistive thermal model, which is computationally expensive since the value of I_{ik} needs to be updated frequently to capture the effect by different T-via distributions. Here, we use a simple way as [14] to calculate I_{ik} by

$$I_{ik}=\sum_{j}H_{ijk}$$

where H_{ijk} is the heat flow from tile t_{ij} on the ith layer to tile $t_{i-1,k}$ on the $i-1$th layer. It is inversely proportional to the thermal resistance between these two tiles. The thermal resistance of this path could be calculated adding the resistance on the path together

$$R_{ijk}=R_{ik}+R_{\mathrm{h}}l_{jk} \tag{9}$$

where R_{h} is the horizontal thermal resistance between two adjacent tiles. l_{jk} is the horizontal Euclidean distance between t_{ij} and t_{ik}.

In the resistive network, there are many heat dissipating paths from a tile t_{ij} to the tiles at the $i-1$th layer. Specially, there are multiple paths connecting tile t_{ij} and $t_{i-1,k}$. It is difficult to calculate the effective thermal resistance between them. In [14], it is found that to consider all these paths would not improve the final results significantly since most of the heat generated by tile t_{ij} flows horizontally to its neighboring tiles then to the $i-1$th layer. In our implementation, we use a simple heuristic method to solve the problem. The thermal resistance calculation

in (9) only considers one heat flow path: horizontally from tile t_{ij} to tile t_{ik}, then vertically from tile t_{ik} to $t_{i-1,k}$. Experimental results of both [14] and our algorithm show that this heuristic method had only a little impact on solution quality.

Thus, the heat flow between two tiles can be calculated by

$$H_{ijk} = (P_{ij} + I_{i+1,j}) \frac{1}{R_{ijk}} \Big/ \sum_{l} \frac{1}{R_{ijl}}$$

where P_{ij} is the heat flow generated by the heat resource in tile t_{ij}. $I_{i+1,j}$ is the vertical heat flow from $i+1$th layer and ith layer.

Using (7)–(9), the intralayer TVP problem on the ith layer could be expressed as

$$\begin{aligned} &\min \sum_{k=1}^{N} m_{ik} \\ &\text{s.t. (a)} \quad R_{ik} \sum_{j=1}^{N} (P_{ij} + I_{i+1,j}) \frac{1/R_{ijk}}{\sum_{l=1}^{N} 1/R_{ijl}} \le \Delta T_i \\ &\qquad k = 1, \ldots, N \\ &\quad \text{(b)} \quad \sum_{k=1}^{N} m_{ik} \ge N m_i \\ &\quad \text{(c)} \quad 0 \le m_{ik} \le C_{ik}. \end{aligned} \tag{10}$$

Constraints (a) specify the maximal thermal gradient of these tiles on the ith layer. If constraints (a) are satisfied for all these device layers, the maximal temperature on the top layer of 3-D chip is below the temperature threshold. Constraint (b) is added since the interlayer via planning only determines the minimal number of T-vias at each layer, which corresponds to an ideal uniform thermal distribution on each layer without T-via insertion. However, if: 1) temperature distribution before T-via insertion is not uniform and 2) hot areas have been occupied by large blocks, more T-vias are needed to be inserted around the hot areas so the total number of T-vias may be larger than the T-via budget calculated by (6). For efficiency reasons, we also ignore the T-via capacity constraints (c) in following derivations, which means all the T-vias can be inserted to desired tiles. The capacity constraints are to be handled when integrating TVP into 3-D floorplanning process.

It is inefficient to solve the constrained NLP directly since constraints (a) are highly nonlinear and the nonlinear terms are as many as the linear terms. It is difficult to deal with the constrained highly nonlinear problem efficiently. However, the NLP could be transformed into a CP by some relaxation and simplification. Below, we introduce the simplified problem formulation and give out a nearly optimal solution for it.

First, we relax the thermal gradient constraints by adding the N constraints for each tile together

$$\sum_{k=1}^{N} R_{ik} \sum_{j=1}^{N} (P_{ij} + I_{i+1,j}) \frac{1/R_{ijk}}{\sum_{l=1}^{N} 1/R_{ijl}} \le N \Delta T_i.$$

By satisfying the relaxed constraint, the average temperature gradient would be upper bounded by the desired maximal temperature gradient between adjacent device layers. This simplification approximates the maximal temperature gradient inside each tile to the average temperature gradient after T-via insertion. It may result in maximal temperature higher than the temperature constraint after T-via insertion on initial 3-D floorplanning result. However, experimental results show that the difference between the maximal temperature and the average temperature inside each layer are within 5 °C after intralayer TVP. The simplification only leads to a little degradation on the solution quality. Moreover, we can use an incremental T-via insertion technique to bring down the maximal temperature, which will be discussed in Section V-D. Thus, it is valid to apply the new constraint for intralayer TVP problem.

From (8), we have

$$\sum_{l=1}^{N} 1/R_{ijl} = \frac{\sum_{l=1}^{N} m_{il} K_{\text{via}} + \left(1 - \sum_{l=1}^{N} m_{il}\right) K_{\text{avg}}}{\alpha (L_{\text{layer}} + L_{\text{interlayer}})/A_{\text{t}}} = \frac{1}{R_i}. \tag{11}$$

Using (9) and (11), the new thermal gradient constraint can be rewritten as

$$\frac{1}{R_i} \sum_{k=1}^{N} \sum_{j=1}^{N} \frac{P_{ij} + I_{i+1,j}}{1 + R_{\text{h}} l_{jk}/R_{ik}} \le N \Delta T_i.$$

The LHS of the above equation is a function of T-via densities inside each tile. Let

$$G(m_{i1}, \ldots, m_{iN}) = \frac{1}{R_i} \times \sum_{k=1}^{N} \sum_{j=1}^{N} \frac{P_{ij} + I_{i+1,j}}{1 + R_{\text{h}} l_{jk}/R_{ik}}.$$

Lemma 2: Function G is convex on the interval [0, 1].

Proof: The partial derivative of G is

$$\frac{\partial G}{\partial m_{ik}} = -\frac{1}{R_i} \times \sum_{j=1}^{N} \frac{(P_{ij} + I_{i+1,j}) \left(\frac{R_{\text{h}} l_{jk}}{R_{\text{via}}} - \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} \right)}{\left(1 + \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} + \left(\frac{R_{\text{h}} l_{jk}}{R_{\text{via}}} - \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} \right) m_{ik} \right)^2}$$

where R_{via} is the thermal resistance of each tile if it is entirely occupied by T-vias, R_{avg} is the thermal resistance of each tile without T-vias. They could be calculated out with (2) and (3). It is shown that each partial derivative is only related to variable m_{ik}, so the second derivatives are as follows:

$$\frac{\partial^2 G}{\partial m_{ik}^2} = \frac{1}{R_i} \times \sum_{j=1}^{N} \frac{(P_{ij} + I_{i+1,j}) \left(\frac{R_{\text{h}} l_{jk}}{R_{\text{via}}} - \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} \right)^2}{\left(1 + \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} + \left(\frac{R_{\text{h}} l_{jk}}{R_{\text{via}}} - \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} \right) m_{ik} \right)^3}$$

$$\frac{\partial^2 G}{\partial m_{ik} \partial m_{is}} = 0, \qquad k \ne s.$$

From the above two conditions, the second derivative of G is

$$\nabla^2 G = \begin{bmatrix} \frac{\partial^2 G}{\partial m_{i1}^2} & \cdots & 0 \\ \vdots & \cdots & \vdots \\ 0 & \cdots & \frac{\partial^2 G}{\partial m_{iN}^2} \end{bmatrix}.$$

Note that $K_{\text{via}} \gg K_{\text{avg}}$, thus $R_{\text{via}} \ll R_{\text{avg}}$. It leads to $\partial^2 G/\partial m_{ik}^2 > 0$, so $\nabla^2 G$ is positive definite and G is a convex function.

Note that constraint (b) is linear so the constraints in the NLP become convex. With above simplification and relaxation, the NLP in (10) could be rewritten as

$$\begin{aligned} &\min \sum_{k=1}^{N} m_{ik} \\ &\text{s.t. (a)} \quad \frac{1}{R_i} \sum_{k=1}^{N} \sum_{j=1}^{N} \frac{P_{ij} + I_{i+1,j}}{1 + R_{\text{h}} l_{jk}/R_{ik}} \le N \Delta T_i \\ &\qquad \text{(b)} \quad \sum_{k=1}^{N} m_{ik} \ge N m_i \end{aligned}$$

which becomes a CP. It can be solved directly with the KKT optimal condition. Let

$$g = \sum_{k=1}^{N} m_{ik} + \mu_1 \left(\sum_{k=1}^{N} m_{ik} - N m_i \right) + \mu_2 \left(N \Delta T_i - \frac{1}{R_i} \sum_{k=1}^{N} \sum_{j=1}^{N} \frac{P_{ij} + I_{i+1,j}}{1 + R_{\text{h}} l_{jk}/R_{ik}} \right).$$

The derived function of g is

$$\frac{\partial g}{\partial m_{ik}} = 1 + \mu_1 - \frac{\mu_2}{R_i} \sum_{j=1}^{N} \frac{(P_{ij} + I_{i+1,j}) \left(\frac{R_{\text{h}} l_{jk}}{R_{\text{via}}} - \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} \right)}{\left(1 + \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} + \left(\frac{R_{\text{h}} l_{jk}}{R_{\text{via}}} - \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} \right) m_{ik} \right)^2}.$$

The KKT optimal condition requires

$$\partial g/\partial m_{ik} = 0, \qquad 1 \le k \le N.$$

From these N equalities

$$\sum_{j=1}^{N} \frac{(P_{ij} + I_{i+1,j}) \left(\frac{R_{\text{h}} l_{jk}}{R_{\text{via}}} - \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} \right)}{\left(1 + \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} + \left(\frac{R_{\text{h}} l_{jk}}{R_{\text{via}}} - \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} \right) m_{ik} \right)^2} = \sum_{j=1}^{N} \frac{(P_{ij} + I_{i+1,j}) \left(\frac{R_{\text{h}} l_{js}}{R_{\text{via}}} - \frac{R_{\text{h}} l_{js}}{R_{\text{avg}}} \right)}{\left(1 + \frac{R_{\text{h}} l_{js}}{R_{\text{avg}}} + \left(\frac{R_{\text{h}} l_{js}}{R_{\text{via}}} - \frac{R_{\text{h}} l_{js}}{R_{\text{avg}}} \right) m_{is} \right)^2}, \quad 1 \le k;\ s \le N. \tag{12}$$

The conditions $K_{\text{via}} \gg K_{\text{avg}}$ and $R_{\text{via}} \ll R_{\text{avg}}$ also results in the following two equalities:

$$\begin{aligned} &1) \quad \frac{R_{\text{h}} l_{jk}}{R_{\text{via}}} \gg \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}} \\ &2) \quad \frac{R_{\text{h}} l_{jk}}{R_{\text{via}}} m_{ik} \gg 1 + \frac{R_{\text{h}} l_{jk}}{R_{\text{avg}}}. \end{aligned}$$

The second equation is validated by experimental results. With 1) and 2), (12) could be approximated by

$$\sum_{j=1}^{N} \frac{(P_{ij} + I_{i+1,j})/R_{\text{h}} l_{jk}}{m_{ik}^2} = \sum_{j=1}^{N} \frac{(P_{ij} + I_{i+1,j})/R_{\text{h}} l_{js}}{m_{is}^2}, \qquad 1 \le k;\ s \le N.$$

Therefore, we derive a nearly optimal solution for the intralayer TVP problem in (10)

$$m_{ik} : m_{is} = \sqrt{\sum_{j=1}^{N} \frac{P_{ij} + I_{i+1,j}}{l_{jk}}} : \sqrt{\sum_{j=1}^{N} \frac{P_{ij} + I_{i+1,j}}{l_{js}}}, \quad 1 \le k;\ s \le N. \tag{13}$$

The equation indicates that the number of T-vias assigned to each tile is proportional to the square root of total incoming heat flow, which is also the vertical heat flow inside that tile. It could be rewritten as

$$m_{ik} : m_{is} = \sqrt{I_{ik}} : \sqrt{I_{is}}. \tag{14}$$

This equation determines the detailed T-via distribution inside each layer. Given a 3-D floorplanning result, the temperature constrained TVP problem can be solved directly by (6) and (13). To apply (13) for T-via density calculation, the heat flow from each tile to other tiles is inversely proportional to the length of the heat dissipating path, l_{jk}. Its value would be zero for $j = k$, which makes the right-hand side of (13) to be infinite. To solve this problem, the horizontal distance from one tile to itself is set to be 1/4 of the sum of its width and height in our implementation, which means most of the heat generated in some tile on the ith layer flows vertically to the $i-1$th layer directly.

It is also shown that the incoming heat flow of tile t_{ik}, I_{ik}, depends on the heat flow from above layer $I_{i+1,k}$. After determining the T-via distribution at the ith layer, the vertical heat flow inside each tile would be updated to determine the via distribution at the $i-1$th layer. Then, T-via distribution at the $i-1$th layer could be determined. Thus, our intralayer TVP algorithm tackles with each device layer one by one: starting from the top layer and ending at the bottom layer.

V. 3-D Floorplanning With via Planning

In this section, we introduce how to apply these analytical solutions for TVP during 3-D floorplanning while maintaining a low computational complexity.

A. Overview

Previous 3-D floorplanning representations and algorithms can be classified into two categories. One kind of them is to extend exist 2-D representation to a real 3-D structure, such as 3-D slicing tree [29] and sequence triple [30], where representation for the third dimension is added. Under current 3-D technology, blocks are located in a limited number of device layers. The real 3-D representation would have too much redundancy in the adding data structure for the z-axis, which is not efficient in both time and space. The other kind of 3-D floorplanning algorithms represents floorplans of different layers with an array of 2-D representations (2-D array), such as two-layer bounded-sliceline grid (BSG) [7], four-layer sequential pair (SP) [4], and four-layer transitive closure graph (TCG) [5]. With these representations, blocks are moved inside each layer or swapped between different layers for solution perturbations. To overcome the limitations of lacking the relative position information of blocks on different layers, the combined bucket and 2-D array [5] representation based on the 2-D-array TCG is proposed.

However, all these 2-D-array-based representations suffer from a much enlarged solution space due to the multiple device layer structure, which makes the 3-D floorplanning problem much more complex and results in longer running time and/or lower solution quality. For a floorplanning problem with N blocks, a representation for 2-D floorplan may be divided into $N^{k-1}/(k-1)!$ different 2-D arrays that represents a k-layer 3-D floorplan. It means each 2-D floorplan representation corresponds to $N^{k-1}/(k-1)!$ possible k-layer 3-D floorplanning results so the solution space is scaled up by $N^{k-1}/(k-1)!$ times [26]. To exploit so large a solution space makes the 3-D floorplanning very time consuming. On the other hand, the benefit of 3-D integration would also be degraded due to the deficiency of these algorithms. It results in lower solution quality, especially in thermal optimization algorithms that already have a large runtime overhead on thermal profiling during iterative improvement process [5].

Note that we have derived analytical solutions for interlayer and intralayer T-via distributions, respectively. Moreover, it is reported that by solving the interlayer and intralayer via planning subproblems, solution quality is similar to solving original TVP problem directly [14]. Based on this observation, we use a two-stage approach to handle the 3-DFP-TVP problem efficiently. With our method, original problem targets are done sequentially. 1) Before floorplanning, blocks are partitioned into different layers and the interlayer T-via distribution is determined. 2) Then, we generate floorplans for all these layers and determine the intralayer T-via distribution inside each layer in an SA process. Fig. 4 is the design flow of our two-stage approach.

The two-stage 3-D floorplanning flow has been applied for wirelength and thermal optimization [6], [26], [27]. Compared with previous flat 3-D floorplanning algorithm, the two-stage approach could improve total wirelength by 15% to 21% with the same interlayer interconnect via budget [26]. Further experimental results show that to achieve the same wirelength result, the two-stage approach is 5X times faster than the flat algorithm

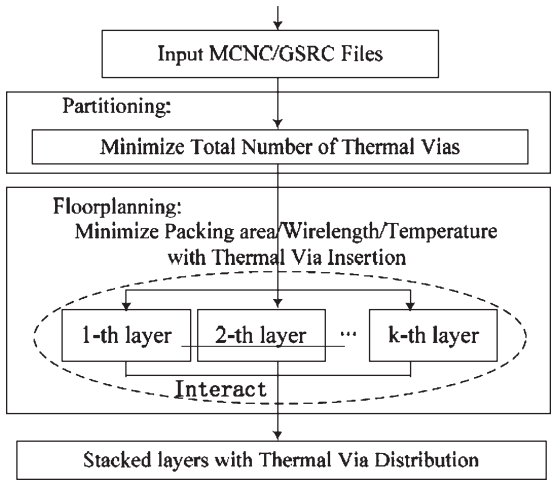


Fig. 4. Design flow of 3-D floorplanner.

[26]. When applied to thermal optimization, it results in much faster convergence speed with similar solution quality [27].

However, this approach brings on another problem: as blocks cannot move to other layers during floorplanning, there would be less possibility to find better 3-DFP-TVP solutions in the much smaller searching space. For example, if we partition blocks into different layers with traditional min-cut methods, which minimize the number of interlayer interconnect vias, it will lead to an increase in the total wirelength in 3-D integration. Experimental results show that interlayer partition with maximal interconnect vias may results in 20% improvement on total wirelength compared with minimal interconnect via partition [30]. The key problem for our 3-D floorplanner is to choose a partition scheme that is most beneficial for T-via reduction so as to scale down the searching space without causing sacrifice on solution quality. Below, we address on this problem and give out an interlayer partitioning solution for T-via number minimization.

B. Interlayer Partitioning

As the positions of blocks remain unknown during interlayer partitioning, the power distribution could not be determined without packing. To estimate total number of T-vias, we assume that the temperature distribution inside each device layer is uniform. The number of T-vias inside each layer can be calculated with (6). To minimize the total number of T-vias, the interlayer partitioning problem is defined as

$$\min \sum_{i=2}^{k} m_i$$

$$\text{s.t.} \sum_{i=1}^{k} P_i = \text{const}$$

$$\frac{1}{\beta} \le \left| \frac{A_i}{A/k} \right| \le \beta, \qquad 1 \le i \le k$$

where A_i is the sum of the areas of blocks that assigned to the ith layer. A is the sum of the areas of all the blocks. β is the area balance ratio factor. In our experiments, it is set to be 1.1.

Based on the analytical solution in (6), this objective can be rewritten as

$$\begin{aligned}\sum_{i=2}^{k} m_i &= \frac{1+\lambda m_2}{\lambda}\sum_{i=2}^{k}\frac{\sqrt{Q_i}}{\sqrt{Q_2}} - k \\ &= \frac{\theta\sqrt{Q_2}\sum_{i=2}^{k}\sqrt{Q_i}}{\lambda\left(T_0 - T_{\text{amb}} - R_{\text{b}}Q_1\right)}\sum_{i=2}^{k}\frac{\sqrt{Q_i}}{\sqrt{Q_2}} - k \\ &= \frac{\theta\left(\sum_{i=2}^{k}\sqrt{\sum_{l=i}^{k}P_l}\right)^2}{\lambda(T_0 - T_{\text{amb}} - R_{\text{b}}\times\text{const})} - k.\end{aligned}$$

With the constraint of area balance between different layers, we assume that the final packing area S to be a constant and estimated by $S = 1.3 \times A/k$. The weighting factor of area usage can also be adjusted according to some experimental results on different testcases. With this assumption, both θ and R_{b} can be calculated out as constants. The interlayer partitioning problem can be rewritten as

$$\begin{aligned}&\min \sum_{l=2}^{k}\sqrt{\sum_{i=l}^{k}P_i} \\ &\text{s.t. } \sum_{i=1}^{k}P_i = \text{const} \\ &\qquad \frac{1}{\beta} \le \left|\frac{A_i}{A/k}\right| \le \beta, \qquad 1 \le i \le k.\end{aligned}$$

This problem can be solved by solving a sequence of subproblems. As P_1 is not included in the objective function, minimizing the objective function corresponds to maximizing P_1. From this observation, we define the subproblem S_1 as

$$\begin{aligned}&\max P_1 \\ &\text{s.t. } \frac{1}{\beta} \le \left|\frac{A_1}{A/k}\right| \le \beta.\end{aligned}$$

This is a classical knapsack problem and can be solved using SA engine or good heuristics. Let P_1^* be the solution for S_1, original interlayer partitioning problem can be redefined as

$$\begin{aligned}&\min \sum_{l=3}^{k}\sqrt{\sum_{i=l}^{k}P_i} + \sqrt{\sum_{i=2}^{k}P_i} = \sum_{l=3}^{k}\sqrt{\sum_{i=l}^{k}P_i} + \sqrt{\text{const}'} \\ &\text{s.t. } \sum_{i=2}^{k}P_i = \text{const} - P_1^* = \text{const}' \\ &\qquad \frac{1}{\beta} \le \left|\frac{A_i}{A/k}\right| \le \beta, \qquad 2 \le i \le k.\end{aligned}$$

It is shown that P_2 is eliminated from the objective function. With the same method, we can define subproblem S_2 and calculate out P_2^*. With a sequence of solving such knapsack subproblems, we get the solution for original interlayer partitioning problem. In our implementation, we use a fast SA engine to solve each knapsack problem as there are only up to several hundred binary variables.

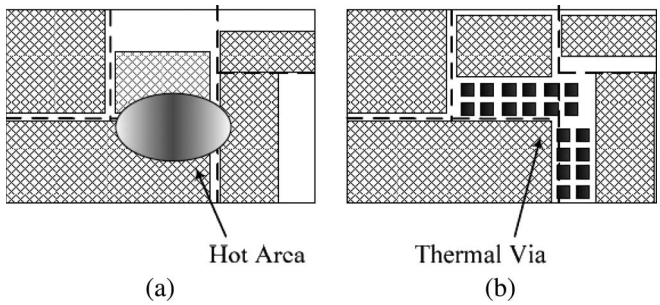


Fig. 5. T-via insertion. (a) No white space budget. (b) White space redistribution.

This solution can be viewed as assigning blocks with higher power density to the bottom layer and blocks with lower power density to top layers. Through this method, heat generated by blocks with higher power density can be dissipated to the heat sink easily. On the other hand, blocks with high power density are not easily to overlap to cause local hot area since they are assigned to the same device layer. The improvement on thermal distribution results in smaller number of T-vias.

C. White Space Redistribution

After assigning blocks to different layers, we generate floorplan on these layers. The intralayer via planning method determines the T-vias assigned to each tile, which should be arranged in white spaces between blocks. Experimental results show that there are many large hot areas occupied by macroblocks and thermal distribution cannot easily be improved since T-vias cannot be inserted directly on these areas as shown in Fig. 5(a). To solve this problem, white space allocation method is needed to change the positions of blocks to achieve enough white space resources for T-via insertion without cause much sacrifice on circuit performance as shown in Fig. 5(b). In the following section, we address on the white space redistribution problem during floorplanning and propose an efficient method to solve it.

In the room-based floorplanning representation, the blocks are packed within the range of the rooms as shown in Fig. 5. To insert T-vias to desired tile as much as possible, the white spaces resource should be allocated as needed. Recently, optimal white space redistribution approach is proposed to minimize total wirelength after floorplanning [19]. The problem is formulated as a linear programming problem and solved by min-cost flow implementation. Although the method guarantees to obtain an optimal wirelength and it is efficient as a postfloorplanning step, it is still too time consuming to be integrated into floorplanning process. With the budget of the T-via insertion, we introduce a novel method to place the blocks in their room to favor via insertion.

Given an m-block set, it divides the chip into at least m rooms and assigns no more than one block to each room. Suppose that the rooms in the ith layer are $\{R_{i1}, \ldots, R_{\text{im}}\}$ and

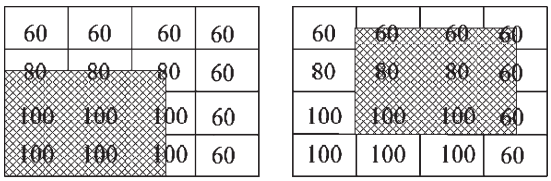


Fig. 6. White space redistribution. (a) $V\_\mathrm{b} = 620$. (b) $V\_\mathrm{b}' = 510$.

$T\_\mathrm{Covered}(R_{il})$ is the set of the tiles covered by the block in R_{il}. The white space ratio in t_{ij} is defined as

$$\mathrm{WS\_ratio}_{ij} = A\_\mathrm{DS}_{ij}/A\_t_{ij}$$

where A_DS_{ij} is the area of white space in tile t_{ij} and A_t_{ij} is the area of t_{ij}. If t_{ij} is covered entirely by the circuit blocks, $\mathrm{WS\_ratio}_{ij} = 0$ and if t_{ij} is not covered by any block at all, $\mathrm{WS\_ratio}_{ij} = 1$. We define the T-via budget in room R_{il} as

$$\nu\_\mathrm{budget}(R_{il}) = \sum_{t_{ij} \in T\_\mathrm{Covered}(R_{il})} m_{ij}(1 - \mathrm{WS\_ratio}_{ij}).$$

In order to optimize T-via insertion, the objective of white space redistribution is to decide the position of blocks in each room so that the T-via budget of the covered tiles is minimized

$$\min \sum_{i=1}^{k} \sum_{l=1}^{m} \nu\_\mathrm{budget}(R_{il}).$$

Note that the desired number of T-vias inserted to t_{ij} is treated as a weighting factor. When minimizing the objective function, more white space resources would be assigned to tiles with larger m_{ij}. Then, more T-vias can be inserted into tiles with more vertical heat flow so that the maximal thermal gradient can be reduced.

As the budgets of T-via insertion are independent in each room, the rooms can be handled one by one. As shown in Fig. 6, we restrict the lower left corner of blocks locating at the lower left corner of tiles. To speed up the computation, the problem is handled in two directions. First, we move the block in the vertical direction to decide the vertical position. Then, we move the block horizontally to fix the horizontal position. Since blocks are restricted in their rooms, the topology and total packing area would not be affected, while total wirelength may be increased slightly.

The white space resources in a floorplan are composed of two parts: 1) dead space within the range of rooms and 2) white space channel between rooms. White space redistribution technique is applied to the dead space within the range of each room. In our implementation, additional white space channels are inserted to occupy about 10% chip area on each testcase; otherwise, it would be too difficult to satisfy the maximal temperature constraint since these macroblocks are too close to each other. Experimental results show that the dead space redistribution technique has only a little impact on final results (5% T-via budget overhead without dead space redistribution) since most of the T-vias are inserted to the white space channels between rooms.

D. 3-D Floorplanning Method

Given a set of blocks, they would be partitioned into different layers with respect to TVP. Then, floorplans of all the layers are generated simultaneously in an SA process with corner block list (CBL) representation [18]. During floorplanning, interlayer T-via distribution is calculated with (6) based on the packing area result. Then, these vias are assigned to the white space inside each tile and the vertical thermal resistance of these tiles would be recalculated when TVP is completed. After T-via insertion, thermal profiling is performed to calculate the maximal on-chip temperature.

The optimality of 3-D floorplanning result is measured by following cost function:

$$\Psi = A + w_1 W + w_2 |T_{\max} - T_0| \tag{15}$$

where A is the packing area of stacked layers. W is the total wirelength. $T_{\max} - T_0$ is the difference between the given temperature threshold and the maximal on-chip temperature after T-via insertion. (w_1, w_2) are the weighting factors. To further speed up the algorithm, the frequency of thermal profiling and TVP is reduced during floorplanning process. Some operations such as changing the orientations of blocks would have little effect on thermal distribution, while swapping two blocks with difference on block dimension or power density may have large effect on final temperature distribution. In our implementation, TVP and thermal profiling are done only after operations with large impacts on thermal distribution, or after several operations without thermal profiling.

There are a few issues worth further discussion.

1) The analytical solutions in (6) and (13) only determine how to proportion the T-vias to each layer and each tile. The precise T-via budget cannot be calculated with these equations since 1) the optimal interlayer T-via budget calculated by (6) is applicable only if the temperature distribution inside each layer is uniform and 2) even if the solution in (13) is close to the optimal intralayer T-via distribution, the maximal temperature constraint may be not satisfied. It depends on the temperature distribution on the bottom layer. If the distribution is uniform, the maximal temperature is below the threshold by satisfying these maximal thermal gradient constraints of adjacent layers; otherwise, the maximal temperature may be higher than the threshold.
2) Even with more T-via budget, the maximal temperature constraint is difficult to be satisfied on some 3-D floorplanning results due to lacking of white space resources at hot areas. The hottest spots are usually occupied by blocks as shown in Fig. 6 and the T-vias could not be inserted directly. This problem may be solved by allocating enough white space for T-via insertion. We can reserve white space channels between any two rooms and

dynamically adjust the width of each width space channel so that all needed T-vias could be inserted. However, the white space channel distribution is determined by T-via distribution, while different white space channel distributions results in different temperature distributions thus affect T-via distribution. It makes the estimation of T-via budget quite difficult to satisfy the temperature constraint and it is too time consuming to be integrated into the SA-based floorplanning process. The runtime inefficiency also prevents previous complicated heat propagation analysis iterations [14] to be integrated to adjust intralayer T-via distribution without white space channel reallocation.

3) In our implementation, T-via budget is calculated with (6) after generating a floorplan and we calculate required white space budgets according to the total number of T-vias. T-via distribution inside each layer is calculated with (13) and a uniform white space channel distribution is used. These techniques avoids above time consuming computation and adjustment on floorplanning result during SA process. However, (6) and (13) are based on the assumption of uniform temperature distributions, the solution in (6) only determine the lower bound for minimal T-via budget on a given floorplanning result. The maximal temperature constraint may not be satisfied due to possible less T-via budget and white space resource at hot areas. After floorplanning, the precise T-via budget is to be increased to satisfy the temperature constraint and the increasing amount is related to floorplanning result and maximal temperature. By using a penalty function on maximal temperature constraint in the objective function (15), we can generate floorplans with more T-vias inserted to hot spots to bring down the maximal temperature. To minimize the objective function during floorplanning results in floorplans with minimal increased T-via budget, which leads to minimal total T-via number, and minimized weighted sum of area and wirelength. This method is beneficial for both performance and thermal optimization. It avoids losing possible good solutions that do not satisfy the strict maximal temperature constraint.
4) If the temperature constraint is not satisfied after floorplanning, the T-via budget should be increased. Suppose $T_{\max}$ is the maximal temperature after TVP. T_0 is the temperature threshold. T_{amb} is the ambient temperature. N_{via} is the T-via budget used during floorplanning and we need ΔN_{via} more T-via budget to bring temperature to meet the constraint. As mentioned in Section IV-B, $R_{\text{via}} \ll R_{\text{avg}}$. Thus

$$T_0 - T_{\text{amb}} \approx PR/(N_{\text{via}} + \Delta N_{\text{via}})$$
$$T_{\max} - T_{\text{amb}} \approx PR/N_{\text{via}}$$

where R is the thermal resistance of one T-via. From the above equations

$$\Delta N_{\text{via}} = \frac{T_{\max} - T_0}{T_0 - T_{\text{amb}}} N_{\text{via}}.$$

TABLE II
BENCHMARKS

	Block #	Net #	Via Size	Tile Array
ami33	33	123	10×10	5×5×4
ami49	49	408	10×10	8×8×4
n100	100	885	1×1	10×10×4
n300	200	1585	1×1	12×12×4
n300	300	1893	1×1	16×16×4

TABLE III
ANALYTICAL SOLUTIONS

	Inter-layer Planning	Intra-layer Planning
m-VPPT	$m_i : m_j = I_i : I_j$ *	$m_{ik} : m_{is} = I_{ik} : I_{is}$
m-ADVP	$\frac{m_i + \xi}{m_j + \xi} = \sqrt{I_i} : \sqrt{I_j}$, $\xi = K_{via}/(K_{avg} S)$	$m_{ik} : m_{is} = I_{ik} : I_{is}$
TVP	$\frac{1 + \lambda m_i}{1 + \lambda m_J} = \sqrt{I_i} : \sqrt{I_J}$, $\lambda = K_{via} / K_{avg} - 1$	$m_{ik} : m_{is} = \sqrt{I_{ik}} : \sqrt{I_{is}}$

*I is the vertical heat flow inside each layer/tile.

These T-vias are distributed to each tile with (6) and (13). The dimensions of the white space channels are adjusted according to T-via distribution to get higher area usage.

VI. EXPERIMENTAL RESULTS

We implemented the proposed 3-DFP-TVP algorithm in C++ language on a Sun V880 750-MHz workstation with Sun OS 5.8. To compare with previous TVP algorithms, we also implement a thermal-driven 3-D floorplanning algorithm, which minimize the maximal temperature directly without T-via insertion. Based on floorplanning results, T-vias are inserted and distributed.

We tested these algorithms on Microelectronics Center of North Carolina (MCNC) benchmarks and Gigascale Systems Research Center (GSRC) benchmarks [14] with four stacked layer integration. Table II shows the block number, net number, via pitch size and tile stack array used for each testcase. In [14], the block sizes in GSRC benchmark are scaled up by ten times so the via pitch sizes are scaled with the same factor in this paper. The power dissipations of each block are ranging from 10^5 to 10^7 W/m^2. The ambient temperature is set to be 27 °C. The maximal temperature threshold is set to be 77 °C (350 in absolute temperature). The bulk substrate thickness is set to be 500 μm, the layer thickness is set to be 20 μm and interlayer thickness is set to be 2 μm. The thermal conductivities of layer and interlayer without T-vias are set to be 119 and 1.1 W/mC, respectively. The thermal conductivity of the silicon in the bulk substrate was set to 150 W/mC and the T-vias were assumed to be copper with a thermal conductivity of 300 W/mC. The chip area is calculated as the product of the maximum width and the maximum height of all device layers. The net length after floorplanning is estimated with half perimeter bounding

TABLE IV
T-VIA BUDGET

	m-ADVP		*m-VPPT*		*TVP*		*TVP-1*	
	T_{max}	T-via#	T_{max}	T-via#	T_{max}	T-via#	T_{max}	T-via#
ami33	76.8	1109	76.7	1360	77.5	981	77.4	995
ami49	77.0	21668	77.1	28793	77.2	19857	77.2	20310
n100	77.2	16731	76.9	25205	77.0	14236	77.1	14533
n200	77.1	14273	76.4	17552	77.1	12566	76.9	12869
n300	76.8	19337	76.5	25995	76.9	17853	76.9	18614
Avg.		1.12		1.51		1.00		1.02

TABLE V
IMPACT OF TVP

	3D Floorplanning + TVP					*3DFP-TVP*				
	T_{max}	T-via#	L_{total}	Area	Cpu(s)	T_{max}	T-via#	L_{total}	Area	Cpu(s)
ami33	77.7	1174	28616	5.54E+05	234	78.2	1025	29174	5.36E+05	275
ami49	78.9	24853	513984	1.72E+07	486	78.6	21966	510179	1.68E+07	587
n100	76.5	17695	89402	7.25E+04	3643	77.2	15167	87523	7.12E+04	4655
n200	77.4	15120	186274	7.42E+04	6394	77.3	13241	183519	7.07E+04	8758
n300	77.6	23655	296175	12.2E+04	12927	77.1	19428	296854	11.8E+04	16311
Avg.		1.16	1.00	1.03	0.79		1	1	1	1

box model and the final temperature profiles on each layer are generated by a detailed FDM simulator.

A. Impact of Analytical Solution

Experimental results in [14] show that m-ADVP achieved 69% reduction on T-via budget than their previous work multilevel via planning proportional to temperature (m-VPPT) [12]. The difference between m-ADVP and m-VPPT is that m-ADVP uses different interlayer T-via distribution. Table III lists the analytical solutions for interlayer and intralayer T-via distribution used in m-VPPT, m-ADVP and our TVP algorithm. It is shown that TVP uses similar solution for interlayer TVP and different solution for intralayer TVP.

To understand the impact on total T-via budget by different T-via distributions, we apply the analytical solutions in m-VPPT, m-ADVP, and TVP on initial thermal driven 3-D floorplanning results. As the white space resource can have great impact on T-via insertion result, T-via distributions may be quite different when there is not enough white space to arrange T-vias. To diminish the effect by white space resource, we take away the capacity constraint of each tile. We insert white space channel between blocks and assume all the T-vias can be inserted to desired tiles. The effect of macroblock obstacles would be discussed in the following section.

Table IV shows the T-via budgets on different testcases. It is shown that the analytical solution in TVP results in 12% and 51% reduction on total T-via number than m-ADVP and m-VPPT, respectively. The reduction in T-via budget compared with m-VPPT benefits from the analytical solution for both interlayer and intralayer TVP. The reduction in T-via budget compared with m-ADVP benefits from the different analytical solution for intralayer TVP. For the same temperature constrained TVP problem, our analytical solution results in better detailed T-via distribution.

Note that our intralayer TVP is based on solving the simplified CP. To apply the solution in (14) directly for TVP, thermal profiling is needed to calculate the vertical heat flow inside each tile. Then, T-vias are inserted and another thermal profiling is done to calculate the temperature distribution after T-via insertion. To enhance the runtime efficiency, we use the solution in (13) to calculate the detailed T-via distribution without explicit thermal profiling. Table IV also lists the result of TVP-1 using intralayer planning solution in (13). It is shown that the approximation (13) is very close to the straightforward solution with explicit thermal analysis. During floorplanning, we apply (13) for intralayer TVP. It results in 18% runtime reduction compared to the explicit thermal profiling approach in [28].

B. Impact of TVP

As mentioned in Section V, T-via distribution may be greatly affected by the macroblock obstacles. Generally, T-vias cannot insert directly at the hottest spots since they are occupied by large macroblocks. They are to be arranged in the white spaces around these blocks. More T-via budget is needed to bring down the maximal temperature to the user specified threshold due to the effect of additional thermal resistance between heat resources and T-vias. An integrated 3-DFP-TVP process may generate floorplans more feasible for T-via insertion. Table V shows the experimental results of the integrated algorithm 3-DFP-TVP and postfloorplanning optimization algorithm TVP. The listed results are floorplanning results after white space channel insertion and adjustment for both 3-DFP-TVP and TVP. Given a set of blocks, the maximal temperature and T-via budget are related to the packing area. As we are not doing fixed-outline 3-D floorplanning, more white space budget results in sacrifice on packing area, but benefits temperature optimization and T-via number minimization. To make fair comparisons, we assign the same white space channel to each testcase for 3-DFP-TVP and TVP. When the runtime over heads range from 18% to 23%, 3-DFP-TVP can reduce T-vias by 16% for the same temperature constraint. The packing area is also reduced by 3% since our method generates floorplan more feasible for T-via insertion. Our method is proven to be efficient and effective for integrating TVP into 3-D floorplanning process with moderate runtime overhead.

VII. Conclusion

We have investigated TVP during the 3-D floorplanning stage. First, we derive a nearly optimal analytical solution for detailed T-via distribution by solving simplified interlayer and intralayer TVP subproblems. Based on the analytical solution, we propose a two-stage approach to integrate dynamic TVP into thermal-oriented 3-D floorplanning process. The novel approach scales down the much enlarged solution space due to multiple device layer structure. Experimental results show that for the postfloorplanning TVP problem, our analytical solution results in fewer T-via budget compared with previous algorithms. Moreover, integrating TVP and floorplanning together can reduce T-vias significantly and improve the area usage moderately compared with the postfloorplanning optimization method.

Our interlayer and intralayer T-via distribution algorithm could also be applied directly for postplacement TVP where T-via is modeled as located in some regular T-via regions [15]. In this modeling, TVP problem becomes much easier. The solution in (13) for the intralayer T-via distribution problem could be used directly to determine the T-via density in each thermal region and it is unnecessary to perform white space redistribution.

Note that our algorithm only focuses on through-the-silicon T-via distribution. The positions of through-the-silicon signal vias could also be optimized for wirelength minimization. It would provide a more accurate estimation result on total wirelength and the distributions for T-vias are also to be affected when we take these signal vias into account. Moreover, routing congestion is not modeled explicitly in our method. Further work would be focused on the congestion and delay optimization with both thermal and signal via planning. It would provide a more accurate estimation on the performance and reliability of the 3-D circuits.

References

[1] S. F. Al-Sarawi, D. Abbott, and P. D. Franzon, "A review of 3-D packaging technology," *IEEE Trans. Compon., Packag. Manuf. Technol. B*, vol. 21, no. 1, pp. 2–14, Feb. 1998.

[2] R. Zhang, K. Roy, C. K. Koh, and D. B. Janes, "Stochastic interconnect modeling, power trends, and performance characterization of 3-D circuits," *IEEE Trans. Electron Devices*, vol. 48, no. 4, pp. 638–652, Apr. 2001.

[3] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3D-ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems on chip integration," *Proc. IEEE*, vol. 89, no. 5, pp. 602–633, May 2001.

[4] P. H. Shiu, R. Ravichandran, S. Easwar, and S. K. Lim, "Multi-layer floorplanning for reliable sytem-on-package," in *Proc. Int. Symp. Circuits Syst.*, May 2004, vol. 5, pp. V-69–V-72.

[5] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs," in *Proc. Int. Conf. Comput.-Aided Des.*, Nov. 2004, pp. 306–313.

[6] Z. Li, X. Hong, Q. Zhou, Y. Cai, J. Bian, H. Yang, P. Saxena, and V. Pitchumani, "A divide-and-conquer 2.5-D floorplanning algorithm based on statistical wirelength estimation," in *Proc. Int. Symp. Circuits Syst.*, May 2005, pp. 6230–6233.

[7] Y. Deng and W. P. Maly, "Interconnect characteristics of 2.5-D system integration scheme," in *Proc. Int. Symp. Phys. Des.*, Apr. 2001, pp. 341–345.

[8] B. Goplen and S. Sapatnekar, "Efficient thermal placement of standard cells in 3D ICs using a force directed approach," in *Proc. Int. Conf. Comput.-Aided Des.*, Nov. 2003, pp. 86–90.

[9] S. Das, "Design automation and analysis of three dimensional integrated circuits," Ph.D. dissertation, MIT, Cambridge, MA, 2004.

[10] S. Das, A. Fan, K. N. Chen, C. S. Tan, N. Checka, and R. Reif, "Technology, performance, and computer aided design of three dimensional integrated circuits," in *Proc. Int. Symp. Phys. Des.*, Apr. 2001, pp. 108–115.

[11] K. Balakrishnan, V. Nanda, S. Easwar, and S. K. Lim, "Wire congestion and thermal aware 3D global placement," in *Proc. Asia South Pacific Des. Autom. Conf.*, Jan. 2005, pp. 1131–1134.

[12] J. Cong and Y. Zhang, "Thermal-driven multilevel routing for 3-D ICs," in *Proc. Asia South Pacific Des. Autom. Conf.*, Jan. 2005, pp. 121–126.

[13] B. Goplen and S. Sapatnekar, "Thermal via placement in 3D ICs," in *Proc. Int. Symp. Phys. Des.*, Apr. 2005, pp. 167–174.

[14] J. Cong and Y. Zhang, "Thermal via planning for 3-D ICs," in *Proc. Int. Conf. Comput.-Aided Des.*, Nov. 2005, pp. 745–752.

[15] T. Zhang, Y. Zhan, and S. S. Sapatnekar, "Temperature-aware routing in 3D ICs," in *Proc. Asia South Pacific Des. Autom. Conf.*, Jan. 2006, pp. 309–314.

[16] T. Y. Chiang, S. J. Souri, C. O. Chui, and K. C. Saraswat, "Thermal analysis of heterogeneous 3D ICs with various integration scenarios," in *Proc. Int. Electron Devices Meeting*, 2001, pp. 681–684.

[17] T. Y. Chiang, K. Banerjee, and K. C. Saraswat, "Compact modeling and SPICE-based simulation for electrothermal analysis of multilevel ULSI interconnects," in *Proc. Int. Conf. Comput.-Aided Des.*, Nov. 2001, pp. 165–172.

[18] X. Hong, S. Dong, G. Huang, Y. Cai, C. K. Cheng, and J. Gu, "Corner block list representation and its application to floorplan optimization," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 51, no. 5, pp. 228–233, May 2004.

[19] X. Tang, R. Tian, and D. F. Wong, "Optimal redistribution of white space for wire length minimization," in *Proc. Asia South Pacific Des. Autom. Conf.*, Jan. 2005, pp. 412–417.

[20] P. Wilkerson, A. Raman, and M. Turowski, "Fast, automated thermal simulation of three-dimensional integrated circuits," in *Proc. Intersociety Conf. Thermal and Thermomechanical Phenomena Electron. Syst.*, Jun. 2004, pp. 706–713.

[21] P. Wilkerson, M. Furmanczyk, and M. Turowski, "Compact thermal modeling analysis for 3D integration circuits," in *Proc. Int. Conf. Mixed Des. Integr. Circuits and Syst.*, 2004, pp. 24–26.

[22] C. H. Tsai and S. M. Kang, "Cell-level placement for improving substrate thermal distribution," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 19, no. 2, pp. 253–266, Feb. 2000.

[23] T.-Y. Wang, Y.-M. Lee, and C. C.-P. Chen, "3D-thermal-ADI: Efficient chip-level transient thermal simulator," in *Proc. Int. Symp. Phys. Des.*, Apr. 2003, pp. 10–17.

[24] P. Li, T. Pileggi, M. Asheghi, and R. Chandra, "Efficient full-chip thermal modeling and analysis," in *Proc. Int. Conf. Comput.-Aided Des.*, Nov. 2004, pp. 319–326.

[25] Y. K. Cheng and S. M. Kang, "An efficient method for hot-spot identification in ULSI circuits," in *Proc. Int. Conf. Comput.-Aided Des.*, Nov. 1999, pp. 124–127.

[26] ——, "Hierarchical 3D floorplanning algorithm for wirelength optimization," *IEEE Trans. Circuits Syst. I, Reg. Papers,* submitted for publication.

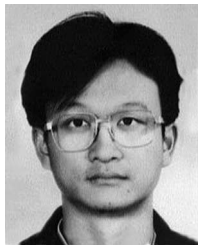
[27] ——"Efficient thermal-oriented 3D floorplanning and thermal via planning for two-stacked-die integration," *ACM Trans. Des. Automat. Electron. Syst.*, vol. 11, no. 2, pp. 325–345, Apr. 2006.

[28] ——, "Integrating dynamic thermal via planning with 3D floorplanning," in *Proc. Int. Symp. Phys. Des.*, Apr. 2006, pp. 178–185.

[29] L. Cheng, L. Deng, and D. F. Wong, "Floorplanning for 3D VLSI design," in *Proc. Asia South Pacific Des. Autom. Conf.*, Jan. 2005, pp. 405–411.

[30] H. Yamazaki, K. Sakanushi, S. Nakatake, and Y. Kajitani, "The 3D-packing by meta data structure and packing heuristics," *IEICE Trans. Fundam.*, vol. E38-A, no. 4, pp. 639–645, Apr. 2000.

[31] S. Das, A. Chandrakasan, and R. Reif, "Design tools for 3-D integrated circuits," in *Proc. Asia South Pacific Des. Autom. Conf.*, Jan. 2003, pp. 53–58.



Zhuoyuan Li received the B.S. and M.S. degrees in computer science from Tsinghua University, Beijing, China, in 2001 and 2003, respectively, where he is currently working toward the Ph.D. degree.

His research interests include incremental physical design, 3-D floorplanning and placement, and thermal analysis and optimization.

Xianlong Hong (M'95–SM'95–F'03) received the B.S. degree from Tsinghua University, Beijing, China, in 1964.

Since 1988, he has been a Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include very large scale integration (VLSI) layout algorithms and design automation systems.

Prof. Hong has served as a Steering Member of the Asia and South Pacific Design Automation Conference (ASPDAC) and the Cochair of the Technical Program Committee of ASPDAC in 1999, 2004, and 2005. He has been an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I since 2002.

Qiang Zhou received the B.S. degree in computer science from the University of Science and Technology of China, Beijing, China, in 1983, the M.S. degree in computer science from Tsinghua University, Beijing, China, in 1986, and the Ph.D. degree from Chinese University of Mining and Technology, Beijing, China, in 2002.

He is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University. His research interests include VLSI layout algorithms and systems.

Shan Zeng received the B.S. degree in computer science from Electronics Science and Technology University, Chengdu, China, in 2004. She is currently working toward the Ph.D. degree in computer science at Tsinghua University, Beijing, China.

Her research interests include interconnect parasitic extraction and circuit simulation.

Jinian Bian received the B.S. degree in computer science from Tsinghua University, Beijing, China, in 1970.

He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University. His research interests include hardware/software codesign and verification, architecture synthesis, interface synthesis, and high-level synthesis.

Wenjian Yu (S'01–M'04) received the B.S. and Ph.D. degrees in computer science from Tsinghua University, Beijing, China, in 1999 and 2003, respectively.

Since August 2003, he has been a Research Assistant with the Department of Computer Science and Technology, Tsinghua University. He was a Postdoctoral Researcher with the University of California, San Diego (UCSD), from September 2005 to February 2006. His research interests include parasitic parameter extraction of interconnects in VLSI circuits, direct boundary-element analysis of electromagnetic fields, and modeling and simulation of VLSI interconnects.

Dr. Yu is a Technical Program Subcommittee member of the Association for Computing Machinery/IEEE Asia South-Pacific Design Automation Conference in 2005 and 2007. He was the recipient of the Microsoft Fellowship Award in 2002 and the distinguished Ph.D. Award from Tsinghua University in 2003. He has served as a reviewer for the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS and the IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES.

Hannah H. Yang received the B.S. degree from Peking University, Beijing, China, and the M.S. and Ph.D. degrees from the University of Texas at Austin, all in computer science.

Since 1995, she has been with the Strategic Computer-Aided Design (CAD) Laboratory at Intel Corporation, Hillsboro, OR, focusing on VLSI physical design algorithms, and integrated logical and physical planning for microarchitecture design exploration. She holds a U.S. patent on distributed logic verification technology called Shark.

Vijay Pitchumani received the B.Tech. degree in electronics from Indian Institute of Technology, Madras, India, and the M.S. and Ph.D. degrees in electrical engineering from Syracuse University, Syracuse, NY.

He is currently a Principal Engineer with Intel Corporation, Santa Clara, CA.

Chung-Kuan Cheng (S'82–M'84–SM'95–F'00) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1976 and 1978, respectively and the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1984.

From 1984 to 1986, he was a Senior CAD Engineer with Advanced Micro Devices, Inc. In 1986, he joined the University of California, San Diego (UCSD), where he is a Professor with the Computer Science and Engineering Department and an Adjunct Professor with the Electrical and Computer Engineering Department. He served as a Chief Scientist with Mentor Graphics in 1999. His research interests include network optimization and design automation of microelectronic circuits.

Prof. Cheng was an Associate Editor of the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS from 1994 to 2003. He was a recipient of the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN Best Paper Awards in 1997 and 2002, and the NCR Excellence in Teaching Award, School of Engineering, UCSD, 1991.