LP Based White Space Redistribution for Thermal Via Planning and Performance Optimization in 3D ICs*

Xin Li* Yuchun Ma* Xianlong Hong* Sheqin Dong* Jason Cong**

*Department of Computer Science & Technology, Tsinghua University, Beijing 100084, P.R.China
Email: {myc, hxl-dcs}@mail.tsinghua.edu.cn

**Department of Computer Science & Technology, UCLA, USA

Abstract: Thermal issue is a critical challenge in 3D IC circuit design. Incorporating thermal vias into 3D IC is a promising way to mitigate thermal issues by lowering down the thermal resistances between device layers. However, it is usually difficult to get enough space at target regions to insert thermal vias. In this paper, we propose a novel analytical algorithm to re-allocate white space for 3D ICs to facilitate via insertion. Experimental results show that after reallocating whitespaces, thermal vias and total wirelength could be reduced by 14% and by 2%, respectively. It also shows that whitespace distribution with via planning alone will degrade performance by 9% while performance-aware via planning method can reduce thermal via number by 60% and the performance is kept nearly unchanged.

I. INTRODUCTION

Three-dimensional (3D) integration has recently drawn much attention due to its potential for reducing the interconnect delay. However, there are some significant challenges along with its adoption and further development. One extremely important issue in 3D IC design is the thermal problem coming from the higher power density and low thermal conductivity.

During the past a few years, several works on thermal optimization have been proposed including thermal-driven floorplanning[1,2], placement[3,4] and routing[5]. Unfortunately, even with complicated thermal-aware approach to improve heat dissipation, the maximum on-chip temperature is still too high for the circuit to operate properly[1]. Introducing thermal vias into the circuit(Fig.1), in fact, is an efficient way to reduce the chip temperature to a satisfactory level[7]. But it is not free to use thermal vias in 3D designs. On one hand, it is costly to fabricate thermal vias and the common thermal via pitch is very large compared that of regular metal wires. On the other hand, thermal vias are usually inserted into white space between blocks, which may block the routing and cause serious congestion problem. Consequently, the number of thermal vias should be reduced as much as possible while temperature constraints must be satisfied as well.

Figure 1 thermal via in 3D IC stack

To optimize the distribution of thermal vias, some previous works[6, 8] formulate and solve the thermal via problem as a post-floorplan procedure to improve heat resource distribution while [9] incorporate thermal via planning into the thermal-driven floorplanning.

Nevertheless, the thermal via distribution is determined by the thermal distribution. Generally speaking, the hotter the region is, the more vias are needed. In fact, hot areas are usually occupied by macro blocks or cells and thermal distribution can not easily be improved since thermal vias can not be inserted directly into these regions as shown in Figure 2(a). As a result, the maximal temperature could not be reduced to desired threshold unless the initial floorplan is modified to facilitate via insertion, which would cause degradation on total wirelength and overall packing area. As shown in Figure 2(b), after modifying positions of the blocks, it could increase total wirelength. In addition, it may even worsen performance by increasing delays along critical paths.

Facing this case, [6] and [8] just insert thermal vias into nearby white spaces or remote regions, which may need much more thermal vias or even may fail to bring down maximal temperature to desired threshold. In this paper, we propose a novel thermal-aware algorithm to redistribute white space for 3D ICs to favor the thermal via insertion while the performance of the design is simultaneously under control.

II. PROBLEM STATEMENT

Thermal-aware whitespace redistribution in 3D IC can be described as: Given a multi-layer packing with a set of $n$ blocks $M={M_1, M_2, \cdots, M_k}$ in $K$ layers, where $w_i$ and $h_i$ specify the dimensions of block $M_i$, respectively, a set of nets $N=\{N_1, N_2, \cdots, N_m\}$ where $N_i, i=1,2,\cdots, m$ describes the connections between blocks, we need to generate a new layout with the same outline of the original packing and the original relative relations between blocks are remained unchanged. But the whitespace between blocks are redistributed so that: 1) the required temperature can be reached with fewer thermal vias; 2) total wirelength is not enlarged evidently or the performance estimation of the design is not degraded seriously.

To deal with the trade-off between multiple objectives, we first estimate thermal via requirement in each tile by a thermal profile model. Then we formulate thermal via requirement and other optimizations into LP models so that multiple objectives and constraints can be handled simultaneously.

III. OVERVIEW OF THERMAL RESISTANCE MODEL

For temperature profiling, we use the same thermal resistive model as [6]. The 3D circuit stack is divided by a two-dimensional array of tile stacks, as shown in Figure 3(a). Each tile stack is

---

*This work is supported by the NSFC 60606007 and Tsinghua Basic Research Fund JC20070021
composed of several vertically-stacked tiles, as shown in Figure 3(b). These tile stacks are connected by lateral thermal resistances, $R_{\text{thermal}}$. Within each tile stack, a thermal resistor $R_i$ is modeled for the $i$-th device layer, while thermal resistance of the bottom layer and silicon substrate is modeled as $R_{b}$, as shown in Figure 3(c).

A tile stack is modeled as a resistive network. The isothermal bases of room temperature are modeled as a voltage source. A current source is present at every node in the network to represent the heat sources. The tile stacks are connected by lateral resistances. One can spatially discretize the system and solve the following equation to determine the steady-state thermal profile as a function of power profile.

$$T = PA^{-1}$$

(1)

where $A$ is an $N \times N$ sparse thermal conductivity matrix. $T$ and $P$ are $N \times 1$ temperature and power vectors.

Given an $n$-block set, it divides the chip into at least $n$ rooms and assigns no more than one block to each room. Since the tile structure is constructed to compute the thermal profile and via distribution, we first figure out ideal via number for each tile to satisfy temperature threshold. Then we assign via requirements to each room according to the tiles it covers. Assume $(x_i,y_i)$ are the coordinates for lower-left corner of block $i$. As shown in Figure 4, tiles are denoted by shadow rectangles, where $\text{area}_{\text{budget}}$ is defined as the overlapping area of room $i$ and tile $j$, which can be calculated as:

$$\text{area}_{\text{budget}} = \begin{cases} \Delta x \times \Delta y & \text{if room } i \text{ covers tile } j \\ 0 & \text{if not cover} \end{cases}$$

(2)

Figure 4  room $i$ covers tile $j$

where:

$$\Delta x = \min(x_i + w_j, x_i + w_j) - \max(x_i, x_i)$$

$$\Delta y = \min(y_i + h_j, y_i + h_j) - \max(y_i, y_i)$$

(3)

Assume $VN_i$ is ideal via number for tile $j$, $\text{via}_{\text{area}}$ denotes the area of a single thermal via. We distribute the ideal via number in tile $j$ to those rooms that cover tile $j$. The exact percentage of via number allocated to room $i$ is proportional to its $\text{area}_{\text{budget}}$. The distribution procedure is a two-stage approach. The first stage is to determine the number of vias allocated to the room and the second stage will find appropriate boundary of the room to receive these vias. Via requirement $\text{req}_j$ divided out to room $i$ from tile $j$ is calculated by:

$$\text{req}_j = \frac{VN_i \times \text{area}_{\text{budget}}}{\sum \text{area}_{\text{budget}}}$$

(4)

which suggests that the more a room covers the tile, the more via requirement from this tile will be allocated to this room.

After via requirement is distributed to the room (transform ideal via number to via requirement), then we must calculate the required area for thermal via insertion in each room. We attach the area requirement in each room to the corresponding block. Assume $\text{left}_{\text{req}}$, $\text{right}_{\text{req}}$, $\text{top}_{\text{req}}$, and $\text{bottom}_{\text{req}}$ are individual via requirements distributed to the left, the right, and the top and the bottom boundary of room $i$ from tile $j$, respectively. The actual distribution percentage is proportional to the part of the boundary cut by the tile. As can be seen from Figure 4, the left bottom corner of room $i$ is inside the tile $j$, thus the via requirements are distributed to $\text{left}_{\text{req}}$ and $\text{bottom}_{\text{req}}$:

$$\text{left}_{\text{req}} = \frac{\text{area}_{\text{req}} - \Delta y}{\Delta x + \Delta y}$$

$$\text{bottom}_{\text{req}} = \frac{\text{area}_{\text{req}} - \Delta x}{\Delta x + \Delta y}$$

(5)

If the entire tile $j$ is covered by the room $i$, we can distribute via requirement evenly to the four boundaries. Above requirement values in each tile are calculated sequentially. Assume $LA_i$, $RA_i$, $TA_i$, and $BA_i$ are total via requirement attached to the left, the right, the top and the bottom boundary of block $i$ of room $i$, respectively, as can be seen in Figure 5.

![Figure 5](image)

Figure 5 via requirement attached to block $i$

Actually, these variables can be achieved by:

$$LA_i = \sum_{\text{for tiles in block } i} \text{left}_{\text{req}}$$

$$RA_i = \sum_{\text{for tiles in block } i} \text{right}_{\text{req}}$$

$$TA_i = \sum_{\text{for tiles in block } i} \text{top}_{\text{req}}$$

$$BA_i = \sum_{\text{for tiles in block } i} \text{bottom}_{\text{req}}$$

(6)

After via requirements for each block are obtained, on the assumption that these required area are rectangle-shaped we then take them as our optimization objective to achieve as much white space as possible for via insertion.

V. LP BASED WHITE SPACE REDISTRIBUTION

A. LP constraints for the topological relations

Given an initial floorplan, it is easy to represent the topological relations in linear constraints preventing overlapping of any pair of rectangular modules $i$ and $j$. In white space redistribution, blocks would deviate from the original positions. Let $(x_i,y_i)$, $(x_j,y_j)$ denote the variable positions of the lower left corners of module $i$ and $j$. From the existing floorplan using 3D representation such as CBA[1], we can capture the corresponding relative positions of modules, which keep unchanged in the optimization process. To prevent overlapping between original modules $i$ and $j$, one of the following linear inequalities actually holds:

$$x_i + w_i - x_j \leq 0 \quad \text{if } i \text{ left to } j$$

$$x_i - w_i - x_j \leq 0 \quad \text{if } i \text{ is right to } j$$

$$y_i + h_i - y_j \leq 0 \quad \text{if } i \text{ is below } j$$

$$y_i - h_i - y_j \leq 0 \quad \text{if } i \text{ is above } j$$

(7)

Meanwhile, assume $W$ and $H$ are width and height of the original packing, respectively, to keep the initial packing area, additional inequalities for any module $i$ are needed as follows:

$$x_i \leq W$$

$$y_i \leq H$$

(8)
section IV, via requirements are attached to the boundaries of blocks. White spaces between blocks can also be partitioned into rectangles around blocks. As Figure 6 shows, the shadow rectangles represent via requirement of each block and we could redistribute the rectangular white spaces to cover these shadow regions. After the optimization, most of via requirement are satisfied, except TA of block i and RAk of block k due to the fixed-outline constraints.

![Figure 6 optimization with via planning](image)

Let variables la, ra, ta and ba represent the actual rectangular white spaces finally allocated to the left, the right, the top and the bottom boundaries of block i. To meet the thermal via requirement as much as possible, the objective then can be written as:

\[
\min_{a,b} \beta L + (RA - ra) + (TA - ta) + (BA - ba)
\]

(9)

where \( \beta \) is the weight factor for block i. The hotter blocks will be assigned with higher weight values.

Intuitively, above objective is to optimize area-difference which is generally quadratic. But if we consider the following equations:

\[
L_i = L_i + r_i, \quad R_i = R_i + r_i, \\
T_i = T_i + w_i, \quad B_i = B_i + w_i \\
l_a = l_i + r_i, \quad r_a = r_i + r_i, \\
t_a = t_i + w_i, \quad b_a = b_i + w_i
\]

where \( L_i, R_i, T_i \) and \( B_i \) are heights(widths) of \( L_i, R_k, T_A \) and \( B_A \) respectively, as Figure 5 shows, while \( l_a, r_i, t_i \) and \( b_i \) are corresponding actual heights(widths) satisfied finally. Then the initial objective can be rewritten as:

\[
\min_{a,b} \beta \left( L_i - l_i \right) + h_i \left( R_i - r_i \right) + w_i \left( T_i - t_i \right) + \left( B_i - b_i \right) \]  

(11)

It is a linear objective, since \( l_i, r_i, t_i \) and \( b_i \) are variables while all the others are known constants.

Take the white space around blocks into consideration, the relative relation constraints in equations (7) should be revised to be:

\[
x_i + w_i + r_i \leq x_j - l_j \quad \text{if i left to j} \\
x_j + w_i + r_i \leq x_i \quad \text{if i is right to j} \\
y_i + h_i + t_i \leq y_j - b_j \quad \text{if i is below j} \\
y_j + h_i + t_i \leq y_i - b_i \quad \text{if i is above j}
\]

And the fixed-outline constraints should also be re-written as:

\[
x_i \geq l_i, \quad y_i \geq b_i, \quad x_i + w_i + r_i \leq W, \quad y_i + h_i + t_i \leq H
\]

(13)

C. Optimization with wirelength(HPWL)

Here we model HPWL optimization with LP formulation as follows. Assume the pins are at the center of the corresponding blocks and each net j has four variables \( x_{\text{max}}, x_{\text{min}}, y_{\text{max}}, y_{\text{min}} \) representing its four boundary edges of the bounding box, \((x_{\text{pin}}, y_{\text{pin}})\) represents the pin location of block i. Since pins are in the bounding box, following LP constraints can be achieved:

\[
x_{\text{pin}} \geq x_{\text{min}}, \quad x_{\text{pin}} \leq x_{\text{max}} \\
y_{\text{pin}} \geq y_{\text{min}}, \quad y_{\text{pin}} \leq y_{\text{max}} \\
x_{\text{pin}} = x_i + w_i / 2, \quad y_{\text{pin}} = y_i + h_i / 2
\]

(14)

Hence, the total wirelength HPWL could be minimized by:

\[
\min \sum_{j \in N} x_{\text{max}} - x_{\text{pin}} + y_{\text{max}} - y_{\text{pin}}
\]

(15)

Where \( N \) is set of nets of the floorplan.

D. LP formulation of performance in micro architecture

In micro-architecture design, there are many different subsystems which separate paths into groups, such as bypass wires between the intALUs. Unlike the wirelength computation, where the wirelength of each net is summed up, in timing analysis, only the delay of the critical path, which is the longest path in a group, will be counted. The performance of the architecture can be evaluated by the weighted sum of number of cycles along the paths. Though the whitespace redistribution does not affect the packing area, blocks’ movement in the rooms may change the distance between blocks so the delays on nets will be modified accordingly. Assume \( K \) is the delay that signal travels by unit wirelength and \( \Phi \) is the clock period. \((x_{\text{pin}}, y_{\text{pin}})\) represents the location of pin, \( lat \) is intrinsic delay of block i. Assume \( T_p, cycle_p \) are the delay on path, number of cycles on the critical path in group g. So that we have:

\[
T_p = \sum_{i,j} \left( |x_{\text{pin}} - x_{\text{pin}}| + |y_{\text{pin}} - y_{\text{pin}}| \right) * K + \sum_{i,j} lat_i \cdot cycle_p = \max(T_p) / \Phi
\]

(16)

where \((i,j)\) is an edge in path p. In [17,18], performance estimation is formulated as a linear function:

\[
\sum c_g \cdot cycle_p
\]

(17)

Where \( c_g \) is the degradation factor for g. Formula (16) implies that for each path p in group g, following inequality is active:

\[
T_p \leq \max(T_p) / \Phi \cdot cycle_p
\]

(18)

To model absolute term || in LP, we introduce two extra variables \( v_{x_{\text{pin}}} \) and \( v_{y_{\text{pin}}} \) then add four equivalent inequalities:

\[
x_{\text{pin}} - x_{\text{pin}} \leq v_{x_{\text{pin}}} - x_{\text{pin}} + x_{\text{pin}} - \leq v_{x_{\text{pin}}} \\
y_{\text{pin}} - y_{\text{pin}} \leq v_{y_{\text{pin}}} - y_{\text{pin}} + y_{\text{pin}} - \leq v_{y_{\text{pin}}}
\]

(19)

then performance-driven LP objective can be written as:

\[
\min \sum_{g \in h} c_g \cdot cycle_p
\]

(20)

E. Simultaneous optimization of multiple objectives

Together with previous LPs, now we can establish our thermal-aware and performance-driven white space redistribution approach. Multiple objectives, i.e., thermal via planning(TV), total wirelength(WL) and performance(P), can be considered at the same time as follows:

\[
\min \alpha * TV + \beta * WL + \lambda * P
\]

(21)

VI. EXPERIMENTAL RESULTS

We have implemented our white space redistribution approach in C++ programming language, and all experiments are performed on a 1.6GHz PC/Intel. The LP solver we used is [10]. The tile array in each layer is set as 30x30 and the initial temperature profiles are generated by a fast resistance simulator from [6]. All initial floorplans are generated from a SA-based thermal-driven floorplanning algorithm[11,13]. We apply the thermal via planning algorithm m-ADVP from [6] to calculate ideal via number in each tile and to insert the thermal vias. M-ADVP has no explicit white space reallocation for thermal via insertion, which just insert thermal vias into nearby white spaces or remote regions if no enough white spaces exist in the target regions. The required temperature threshold is set to be 77°C.

A. White space redistribution for thermal via planning

Firstly we optimize the white space and total wirelength simultaneously on some GSRC benchmarks. The power dissipations of each block are assigned to the same value with [6] (ranging from 10^8 W/m^2 to 10^9 W/m^2). For each benchmark, we generated two different initial floorplans with different max temperatures. Table 1 shows the experimental results of our approach. The cpu column is runtime of the whitespace redistribution procedure. As can be seen, for the same temperature constraints, our algorithm can reduce both thermal vias by 14% and total wirelength by 2% compared to the original packing results. As an illustration, figure 7 provides the floorplans before and after white space redistribution for thermal via insertion.
In this paper, we have presented a LP based thermal-aware white space redistribution algorithm in 3D ICs. Based on explicit thermal profile, the thermal via requirements are estimated, then attached to blocks and finally solved by our LP approach. We have also shown that performance can be efficiently modeled in this approach. Experimental results show that after reallocating whitespaces, the amount of thermal vias and total wirelength could be reduced by 14% and 2%, respectively. In addition, to show the trade-off between via planning and performance, we take Alpha 21264 microprocessor as a design driver, it shows that whitespace redistribution T-via optimization only T-via + performance optimization

References