This document is downloaded from DR-NTU (https://dr.ntu.edu.sg)
Nanyang Technological University, Singapore.

Thermal via allocation for 3-D ICs considering temporally and spatially variant thermal power

Yu, Hao; Shi, Yiyu; He, Lei; Karnik, Tanay

2008

Yu, H., Shi, Y., He, L., & Karnik, T. (2008). Thermal via allocation for 3D ICs considering temporally and spatially variant thermal power. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 16(12), 1609-1619.

https://hdl.handle.net/10356/92285

https://doi.org/10.1109/TVLSI.2008.2001297

© 2008 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder. http://www.ieee.org/portal/site This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder.

Downloaded on 23 Aug 2022 03:17:05 SGT

Thermal Via Allocation for 3-D ICs Considering Temporally and Spatially Variant Thermal Power

Hao Yu, *Member, IEEE*, Yiyu Shi, Lei He, *Senior Member, IEEE*, and Tanay Karnik, *Senior Member, IEEE*

***Abstract*—The existing 3-D thermal-via allocation methods are based on the steady-state thermal analysis and may lead to excessive number of thermal vias. This paper develops an accurate and efficient thermal-via allocation considering the temporally and spatially variant thermal-power. The transient temperature is calculated by macromodel with a one-time structured and parameterized model reduction, which also generates temperature sensitivity with respect to thermal-via density. The proposed thermal-via allocation minimizes the time-integral of temperature violation, and is solved by a sequential quadratic programming algorithm with use of sensitivities from the macromodel. Compared to the existing method using the steady-state thermal analysis, our method in experiments is 126× faster to obtain temperature, and reduces the number of thermal vias by 2.04× under the same temperature bound.**

***Index Terms*—3-D integrated circuit (IC) design, cooling technology, sequential programming, structured and parameterized macromodel, thermal power management.**

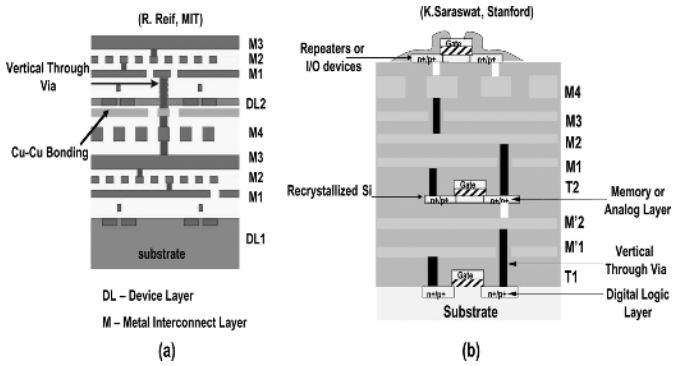

Fig. 1. Typical 3-D IC techniques including: (a) Cu-Cu wafer bonding and (b) crystallization of α-Si. They both have active device layers, inter-layer dielectrics, vertical through vias, and the substrate.

NOMENCLATURE

N	Number of tiles.
K	Number of critical tiles.
p:	number of input ports
q	Order of reduced models.
s/h	Frequency point/time-step.
G_0	nominal thermal conductance state matrix
C_0	Nominal thermal capacitance state matrix.
B/L	Topology matrix describing input/output ports.
A_i	Via density of ith tile.
$\mathbf{A}$	Via density vector of a set of critical tiles.
X	Topology matrix describing where to insert vias.
g/c	Conductance/capacitance of one via with unit area.
x/y	State variable of temperature (at output).
$x^{(0)}/y^{(0)}$	Nominal temperature (at output).
$x^{(1)}/y^{(1)}$	First-order sensitivity (at output).
$x^{(2)}/y^{(2)}$	Second-order sensitivity (at output).
f	Thermal-violation integral.
T_{ceiling}	Targeted ceiling temperature.

Manuscript received February 01, 2007; revised July 03, 2007. Current version published November 19, 2008. This work was supported by the National Science Foundation CCF-0448534 and by UC-MICRO fund from Intel.

H. Yu is now with Berkeley Design Automation, Santa Clara, CA 95054 USA (e-mail: hao.yu@berkeley-da.com).

Y. Shi and L. He are with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: yshi @ee.ucla.edu; lhe@ee.ucla.edu).

T. Karnik is with the Intel Circuit Research Lab., Hillsboro, OR 97124 USA (e-mail: tanay.karnik@intel.com).

Digital Object Identifier 10.1109/TVLSI.2008.2001297

I. INTRODUCTION

THE existing 2-D high-performance system-on-chip (SoC) design is limited by the interconnect delay and device density. 3-D integration [1]–[4] to stack multiple active layered integrated circuits (ICs) is effective to improve the interconnect performance and increase the transistor packing density. Fig. 1 shows the diagram of typical 3-D IC design including the active device layers, vertical through vias, and substrate. Due to the increased power density, heat removal is extremely important in 3-D-ICs [1]. It is well known that excessively high temperature can significantly degrade the reliability of interconnect and device, and cause functional or timing failures through the electrothermal coupling [5]–[13]. The temperature-aware physical design, therefore, becomes important from the early planning stage to the final verification stage [14]–[18].

Because vertical through vias are effective thermal conductors, one effective heat removal approach in 3-D IC is to use vertical through vias to remove heat from stacked silicon layers to the heat-sink that is often on top of the stack. Same as the existing work [16]–[18], we assume that the vertical through vias are aligned through layers and called as *thermal vias*. The 3-D thermal via planning is thereby to allocate vias in order to

1063-8210/$25.00 © 2008 IEEE

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.

alleviate the temperature hotspots at each silicon layer. Since thermal vias consume the routing resource its number ought to be constrained. Assuming a steady-state thermal analysis (based on thermal resistance model), thermal-via allocation has been studied during the placement [16] and routing [17]. The steady-state analysis, however, ignores the temporal and spatial variations of the thermal-power in the modern VLSI design. Due to different workloads and dynamic-power-management techniques such as clock gating, power has both temporal and spatial variations [8], [19]–[21]. A transient thermal-power is thereby defined as the running average of the power at the range of nanosecond over the scale of the thermal time-constant at the range of millisecond [19]. To obtain a solution without the thermal violation, the methods in [16] and [17] need to assume a "steady" maximum thermal-power *simultaneously* for all regions. Because it is rare if not impossible for different regions to simultaneously reach their maximum thermal-power, the methods in [16] and [17] may lead to excessive number of thermal vias.

A cycle-accurate "dynamic" thermal simulator Hotspot [8], [13] has been developed at the micro-architecture level. It is based on a thermal-resistance-capacitance (*RC*) model to calculate the transient temperature. However, they [8], [13], [16], [17] need to directly solve the matrix-formed state equation, and thereby are inefficient to be utilized during the system-level optimization for large scale designs. To achieve a design closure in manageable time, those methods have to sacrifice the physical level detail. Their procedures are either based on oversimplified iterations [16] or based on the approximated square-root-relation [17] between the temperature and thermal-via. They may not converge or lead to inaccurate results. Therefore, it may be beneficial to generate a macromodel that could remain the dominant physical-level detail, yet enable the efficient gradient-search of an optimized system-level design driven by the sensitivity.

In this paper, an accurate yet efficient thermal-via allocation is proposed with consideration of the temporal and spatial variations of the thermal-power. We assume that the signal routing congestion is known a priori, and calculate the transient temperature utilizing macromodel provided by a *structured and parameterized model reduction*, which also generates the temperature sensitivity with respect to the thermal-via density. By defining a *thermal-violation integral* based on the transient temperature, a nonlinear optimization problem is formulated to allocate thermal-vias and minimize thermal violation integral under the signal routing constraint. This optimization problem is transformed into a sequence of quadratic programming (SQP) subproblems using sensitivities provided by the macromodel. Experiments show that compared to the steady-state thermal analysis, our method is 126× faster to obtain the temperature profile, and reduces the number of thermal vias by 2.04× under the same temperature bound.

The rest of this paper is organized as follows. In Section II, we present the preliminary for 3-D thermal model and analysis. In Section III, we formulate a nonlinear optimization for allocating the thermal-via driven by the thermal-violation integral, and propose a sequential programming to efficiently solve the optimization. In Section IV, we discuss a structured and parameterized macromodel to generate the nominal transient temperature and its sensitivities. In Section V, we present the overall algorithm for the thermal-via allocation and experimental results. We conclude in Section VI. The preliminary results of this paper were presented in [22].

TABLE I
THERMAL AND ELECTRICAL DUALITY

Temperature	Voltage state variables ($x(t)$)
Input Thermal-Power	Input Current sources ($u(t)$)
Thermal conductance	Electrical conductance (G)
Thermal capacitance	Electrical capacitance (C)

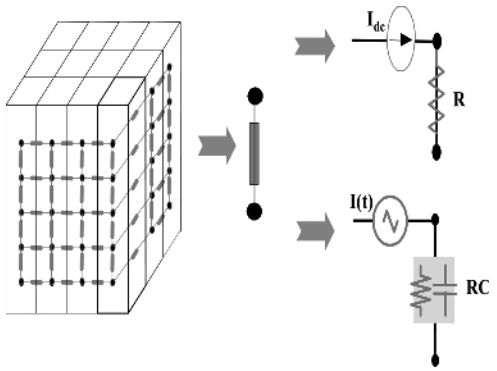

Fig. 2. Thermal models extracted for the steady-state analysis and the transient analysis.

II. 3-D THERMAL MODEL AND ANALYSIS

In this Section, we present how to build a parameterized thermal model for 3-D IC layout, discuss the time-variant thermal power and thermal analysis, and briefly review the existing macromodeling approach.

A. Dynamic and Parameterized Thermal Model

There is a well-known duality between electrical and thermal systems as shown in Table I. As temperature is analogous to voltage, the heat flow can be modeled by a current passing through a pair of thermal-resistance and capacitance driven by the current source, which models the power dissipation. Moreover, because the transient temperature needs to become stable when the steady state is reached after sufficient time, the boundary condition at the chip-surface needs to be specified as the ambient temperature. In this paper, the C4 package is assumed and the packaging and heat-sink are modeled by a simple 1-D resistor network with attached external voltage sources to model the ambient temperature.

Each active device layer and the inter-layer dielectric in 3-D layouts can be uniformly discretized into N tiles by the finite difference method. As shown in Fig. 2, in a steady-state analysis, tiles are connected by the thermal resistance R, and heat-sources are modeled by the time-invariant current source. Then, the steady-state temperature can be obtained by directly solving a time-invariant linear equation. In contrast, as for a transient analysis, tiles are connected by the thermal resistance and capacitance *RC*, and heat sources are modeled by the time-variant current source.

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.

Usually, the granularity of discretized 3-D IC smaller than thermal space constants might not be necessary [13]. Because the proposed method is targeted during the physical design, the granularity of discretized thermal model in this paper can be smaller than those for the microarchitecture design [8], [13]. Moreover, the designs in 3-D IC usually require to consider many heterogeneous components in one system, it also leads to a more complicated thermal model than that for the 2-D IC.

In addition, two types of tiles are specified in this paper: *critical tiles* and *input tiles*. Critical tiles are those tiles with hottest temperatures that can cause thermal violations leading to reliability or timing/functionality failures at those locations. The critical tiles can be pre-characterized during the early design stage, or from an initial full-chip transient simulation. To probe these critical tiles, a topological matrix L (adjacent matrix) can be specified. Input tiles are those tiles with the time-variant power-dissipation (or heat) $u(t)$ averaged at the scale of the thermal time constant. To inject heat at these input ports, a topological matrix B (adjacent matrix) can be specified.

Note that our design parameter here is the thermal-via density. The larger the thermal-via density in one tile, the more heat that can be convected away through layers to the heat sink. An ith tile has a thermal-via area A_i, which is related to the thermal-via density ρ_i by

$$\rho_i = \frac{A_i}{a}$$

where a is the unit area of thermal-via determined by the process technology. Therefore, A_i is used to represent the thermal-via density at ith tile in the sequel. In addition, we assume that thermal vias have a continuous conduct-path from the bottom to the top with alignment.

Then, thermal vias are inserted as follows. An insertion (incident) matrix $X\ (\in R^{N\times N})$ is used to record the location and the number of added vias. If a via is added between two nodes m and n of two vertical-adjacent layers, its insertion matrix is

$$X(k,l) = X(l,k) = \begin{cases} -1, & \text{if } k = m, l = n \\ \sum_l |X(k,l)|, & \text{if } k = l \\ 0, & \text{else.} \end{cases} \quad (1)$$

Accordingly, given the width w and the thickness t of one thermal-via, we have the following topological matrix g_i/c_i for one inserted unit-via conductance/capacitance

$$g_i = \left(\frac{k_1}{t}\right) X_i \qquad c_i = (k_2 t) X_i$$

where k_1 and k_2 are thermal conductive/capacitive constants of the thermal-via.

The *parameterized thermal model* is then constructed as follows. We first define the parameterized state matrices

$$G = G_0 + \sum_{i=1}^{K} A_i g_i \quad C = C_0 + \sum_{i=1}^{K} A_i c_i. \quad (2)$$

Note that G_0 and $C_0\ (\in R^{N\times N})$ are nominal conductive and capacitive matrices of discretized thermal networks, which entry is simply composed of the thermal conductance and capacitance for the 3-D layout without the thermal-via. $\sum_{i=1}^{K} A_i g_i$ and $\sum_{i=1}^{K} A_i c_i$ are the parameterized conductive and capacitive matrices of thermal vias, where via-density A_i is the parameter and K is the number of critical titles.

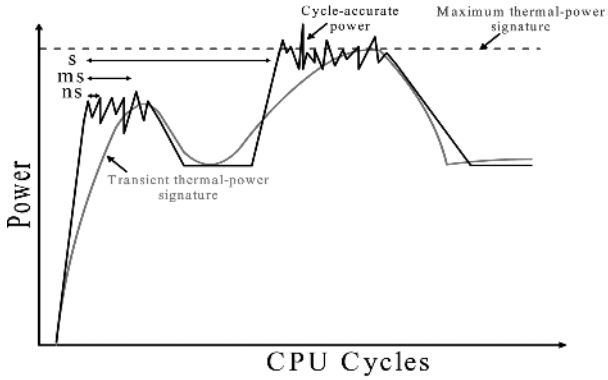

Fig. 3. Definitions of the cycle-accurate power, transient thermal-power, and maximum thermal-power at the different scale of time-constant.

Accordingly, the thermal-RC circuit can be described by the modified nodal analysis (MNA)

$$Gx(\mathbf{A},t) + C\frac{dx(\mathbf{A},t)}{dt} = Bu(t)$$
$$y(\mathbf{A},t) = B^T x(\mathbf{A},t) \quad (3)$$

or in the frequency (s) domain with the Laplace transformation by

$$(G + sC)x(\mathbf{A},s) = Bu(s)$$
$$y(\mathbf{A},s) = L^T x(\mathbf{A},s) \quad (4)$$

where $\mathbf{A} = [A_1, \ldots, A_K]$ is the parameter-vector of thermal-via density, and B and $L\ (\in R^{N\times p})$ are the adjacent matrices to select the input tiles u and critical tiles y. Note that $u\ (\in R^{p\times 1})$ is the current source to model the power input for the thermal network. As defined in [19] (see Fig. 3), a *transient thermal power* is the running average of the cycle-accurate (often in the range of nanoseconds) power over several thermal time constants (often in the range of milliseconds), and a constant *maximum thermal-power* is defined as the maximum of the transient thermal-power. Due to the increasing use of dynamic power managements [8], [19]–[21], the thermal power varies not only spatially but also temporally. The same does the temperature. Moreover, the different applications (workloads) can also lead to the spatial and temporal variation of the on-chip temperature. As a result, the previous via-planning [16], [17] based on the steady-state thermal analysis has to assume the maximum thermal-power simultaneously for all chip regions. This is rare, however, if not impossible for different regions to simultaneously reach their maximum thermal-power. The planned via using steady-state analysis thereby may lead to excessive numbers of vias. This becomes the primary motivation of this paper to study the via-planning problem using a dynamic (or transient) thermal model.

B. Macromodel by Moment Matching

On the other hand, blindly applying the thermal transient analysis is expensive because it is not efficient to solve (4) large-

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.

sized ($N \sim 1\,000\,000$) thermal circuits. Similar to the macromodeling for the interconnect network, the moment-matching based model order reduction can be used to obtain a 3-D IC thermal-macromodel with compact-sized q ($q \ll N$), which not only has a smaller matrix size but also preserves the dominant system response. Below, the principle of macromodeling by the model order reduction is briefly reviewed.

The existing macromodeling approach is mainly based on the subspace-projection [23], [24]. By defining two moment generation matrices expanded at one selected frequency s_0 as

$$\mathcal{A} = (G + s_0 C)^{-1} C \qquad \mathcal{R} = (G + s_0 C)^{-1} B$$

it is easy to verify that the solution of (4) is contained in the subspace spanned by $\mathcal{A}$ and $\mathcal{R}$

$$\text{span}\{\mathcal{R}, \mathcal{A}\mathcal{R}, \ldots, \mathcal{A}^{n-1}\mathcal{R}, \ldots\}.$$

Accordingly, a nth-order block Krylov subspace can be defined by

$$\mathcal{K}(\mathcal{A}, \mathcal{R}, n) = \text{span}\{\mathcal{A}, \mathcal{A}\mathcal{R}, \ldots, \mathcal{A}^{n-1}\mathcal{R}\}$$

where $n = \lfloor q/p \rfloor$.

By applying the Block Arnoldi orthonormalization [24], the spanned subspace by a smaller-dimensioned projection matrix V ($\in R^{N \times q}$) can be found to contain the Krylov subspace

$$\mathcal{K}(\mathcal{A}, \mathcal{R}, n) \subseteq \text{span}\{V\}.$$

Using such a V to project the original state matrices ($R^{N \times N}$), respectively

$$\hat{G} = V^T G V \quad \hat{C} = V^T C V \quad \hat{B} = V^T B \quad \hat{L} = V^T$$

a dimension reduced macromodel ($R^{q \times q}$)

$$\hat{H}(s) = \hat{L}^T (\hat{G} + s\hat{C})^{-1} \hat{B}$$

can be obtained. Note that $\hat{H}$ can accurately approximate the original system H

$$H(s) = L^T (G + sC)^{-1} B$$

by matching the first n block moments expanded at s_0 [23], [24]. Usually, as the time-constant of a thermal *RC* network is much larger than that of an electrical *RLC* network, its dynamic response can be accurately characterized by a few dominant poles using the subspace-projection-based macromodeling. The error of reduced model depends on the selection of the reduced-order q. A detailed analysis of numerical error bound can be found in [23] and [24].

To further obtain the sensitivity information, the parametrized moments [25] can be obtained by expanding (4) at selected parameter points. However, because parameterized moments have coupled frequency and parameter variables, its dimension grows exponentially, preventing practical use. This is improved in [26] by separately expanding moments of parameters from the frequency. It results in an augmented state matrix containing the nominal state and the expanded states, i.e., sensitivities with respect to parameters.

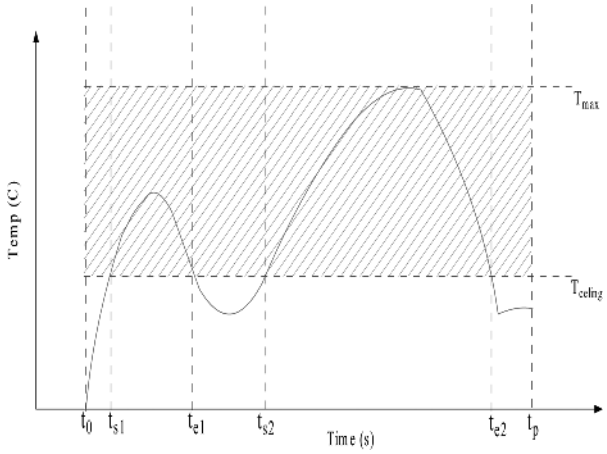

Fig. 4. Figure of merit using thermal-violation integral with defined ceiling temperature under an input of transient thermal-power signature.

Nevertheless, all these approaches [23]–[26] apply a flat projection during the reduction. The reduced state matrices are dense and the reduced state variables have coupled nominal values and sensitivities. It is unknown how to separate parametrized sensitivities from the reduced macromodel, and utilize those sensitivities for the optimization. We will address the solution in Section IV.

III. Thermal-Via Allocation Problem

In this Section, an accurate figure of merit, thermal-violation integral, is first defined to consider the transient thermal integrity. Then, a thermal-via allocation problem is consequently formulated as a nonlinear optimization problem, which is relaxed and solved by a sequence of quadratic programmings with use of sensitivities.

A. Thermal-Violation Integral

A *thermal-violation integral* is the integral of the transient temperature above a user-specified ceiling temperature T_{ceiling}

$$\begin{aligned} f_i(\mathbf{A}) &= \int_{t_0}^{t_p} \max[y(\mathbf{A}, t), T_{\text{ceiling}}] dt \\ &= \int_{t_s}^{t_e} [y(\mathbf{A}, t) - T_{\text{ceiling}}] dt \end{aligned} \tag{5}$$

where t_0 and t_p define the evaluation time-period, which is a sequence that sufficiently contains the possible schedule of the dynamic power management or the given possible workloads. In addition, note that the interval $[t_s, t_e]$ is determined by comparing

$$\max[y(\mathbf{A}, t), T_{\text{ceiling}}]$$

which can contain multiple intervals. Recall that $\mathbf{A}$ is the parameter-vector of the thermal-via density at K critical tiles.

As shown in Fig. 4, the integral is actually the area above the T_{ceiling}. This definition captures the fact that a thermal-violation occurs only when the transient temperature is above the temperature bound for a long enough period. A similar merit is used for noise estimation in [27].

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.

The figure of merit for a group of K critical tiles needs also to be defined. Because it is seldom to happen that different critical tile reaches its targeted ceiling temperature simultaneously, a *global* thermal violation integral

$$f_g(\mathbf{A}) = \sum_{i=1}^{K} f_i(\mathbf{A}) \tag{6}$$

is defined in addition to those *local* thermal violation integral f_is. Accordingly, a *thermal violation integral vector* is defined by composing $f_i (i = 1, \ldots, K)$ and f_g

$$\mathbf{f}(\mathbf{A}) = [f_1(\mathbf{A}), \ldots, f_K(\mathbf{A}), f_g(\mathbf{A})]. \tag{7}$$

The thermal-violation integral vector $\mathbf{f}(\mathbf{A})$ is used as an accurate objective function in the sequel to be minimized.

Note that for the steady-state analysis, the input of the maximum thermal-power signature results in a constant maximum temperature $T_{\max}$. Hence, the hotspot reduction by the steady-state solution is equivalent to reduce a rectangular area defined between $T_{\max}$ and T_{ceiling}, obviously an overestimated violation integral (see Fig. 4). It becomes even worse for the total violation integral. The reason is that each critical tile has a different transient thermal-power signature, and hence, their maximum usually does not happen at the same time. As a result, the thermal-violation integral from a transient solution is more accurate to guide the thermal-via allocation than from a steady-state one.

B. Problem Formulation

To minimize the total violation integral, thermal vias are allocated at each pair of adjacent layers. With consideration of the congestion from vertical signal vias, $A_{\max}$ and $(A_i)_{\max}$ are the *total* available space and *local-tile* available space for inserting thermal vias, which are assumed to be provided by the user. Accordingly, a nonlinear optimization problem is formulated as

$$\underline{\text{Problem 1:}}\ \min \mathbf{f}(\mathbf{A})$$

$$\text{s.t.} \sum_{i=1}^{K} A_i \le A_{\max} \tag{8}$$

$$0 \le A_i \le (A_i)_{\max}, \quad (i = 1, \ldots, K). \tag{9}$$

The constraint (8) is a *global constraint* implying that the total thermal-via density is limited by the $A_{\max}$, and the constraint (9) is a *local constraint* implying that the local thermal-via density at ith tile is limited by $(A_i)_{\max}$. Note that $A_{\max}$ is not always the simple summation of $(A_i)_{\max}$. It is decided by not only the total available routing resources, but also other considerations such as the fabrication cost at different regions. In this paper, we assume that $A_{\max}$ and $(A_i)_{\max}$ are provided by designers.

Moreover, the previously mentioned local and global constraints in Problem 1 can be unified into one constraint with use of one topology matrix $\mathbf{U}$ $(\in R^{(K+1)\times(K)})$

$$\mathbf{U} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & 1 \\ 1 & 1 & \ldots & 1 \end{bmatrix}. \tag{10}$$

As a result, we have

$$\mathbf{UA} \le \mathbf{A}_{\max} \tag{11}$$

where $\mathbf{A}_{\max} = [(A_1)_{\max}, (A_2)_{\max}, \ldots, (A_K)_{\max}, A_{\max}]^T$.

Note that Problem 1 becomes semi-definite [27] and can be solved with the provable convergence when the evaluation period t is discretized into sufficiently small intervals [27]. To efficiently solve Problem 1, the Lagrangian relaxation is used below. The constraint function can be added to the objective function using a vector of Lagrangian multiplier $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]$. As a result, the primal problem (Problem 1) has the following dual problem:

$$\underline{\text{Problem 2:}} L(\mathbf{A}, \boldsymbol{\lambda}) = \mathbf{f}(\mathbf{A}) + \boldsymbol{\lambda} \cdot \mathbf{h}(\mathbf{A}) \tag{12}$$

where

$$\mathbf{h}(\mathbf{A}) = \mathbf{UA} - \mathbf{A}_{\max}. \tag{13}$$

The dual problem can be solved by a sequence of subgradient optimization [28] problems as discussed in the following.

C. Sequential Programming

In general, $f_k(\mathbf{A})$ is a nonlinear function with respect to all A_i's $(i, k = 1, \ldots, K)$. However, if the via-density (area) A_i is only changed by a small amount δA_i $\left(\delta A_i = A_i - A_i^{(0)}\right)$ around the nominal value $A_i^{(0)}$, the corresponding change of $f_k(\mathbf{A})$ can be linear or quadratic with respect to δA_i. Therefore, $f_i(\mathbf{A})$ can be approximated with the Taylor's expansion at the nominal values $f_k^{(0)}(\mathbf{A}^{(0)})$ by

$$f_k(\mathbf{A}) \approx f_k^{(0)}(\mathbf{A}_0) + \sum_{i=1}^{K} \frac{\partial f_k}{\partial A_i} \delta A_i + \sum_{i,j=1}^{K} \frac{\partial^2 f_k}{\partial A_i \partial A_j} \delta A_i \delta A_j. \tag{14}$$

$h_k(\mathbf{A})$ can be expanded in a similar fashion.

Therefore, a sequence of subgradient optimization problems can be formulated for Problem 2

$$\underline{\text{Problem 3:}} \min \nabla \mathbf{f}(\mathbf{A})^T \boldsymbol{\delta}\mathbf{A} + \frac{1}{2} \boldsymbol{\delta}\mathbf{A}^T H \boldsymbol{\delta}\mathbf{A} + \lambda \cdot \nabla \mathbf{h}(\mathbf{A}) \boldsymbol{\delta}\mathbf{A}. \tag{15}$$

Note that

$$\nabla \mathbf{f} = \int_0^{t_p} y^{(1)} dt$$

is the first-order sensitivity, and

$$H = \begin{bmatrix} \int_0^{t_p} y_{1,1}^{(2)} dt & \ldots & \int_0^{t_p} y_{1,K}^{(2)} dt \\ \vdots & \ddots & \vdots \\ \int_0^{t_p} y_{K,1}^{(2)} dt & \ldots & \int_0^{t_p} y_{K,K}^{(2)} dt \end{bmatrix}$$

is the Hessian matrix composed by the second-order sensitivity. In addition, $\nabla h = \text{const}$.

At one iteration, the solution from the quadratic programming problem is used as the intermediate solution of the original nonlinear problem. Then, those coefficients $\nabla \mathbf{f}$ and H are updated and employed to form a new quadratic programming at the new nominal values. The optimization terminates when the convergence criterion is achieved. This is called *sequential quadratic programming* (SQP) [28]. Note that the convergence of the SQP depends on the range of the calculated $\delta \mathbf{A}$. The

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.

quadratic programing may not be accurate to approximate the original nonlinear programming if this range is too large. On the other hand, the quadratic programing may converge slowly if this range is too small. As shown in Algorithm 1, a geometric regression procedure [28] is utilized in this paper to select an optimized subgradient. As a result, the range of $\delta\mathbf{A}$ can be properly determined in our experiment, and hence, the sequence of quadratic programs converges in a few iterations under the specified accuracy.

Algorithm 1 Subgradient Optimization using Structured Parameterized Macromodel

Initialize: $(\mathbf{A}_0, \alpha_0, \boldsymbol{\lambda}_0, H_0, k)$;
Solve: $\widetilde{y}_0$ using **(20)**;
Solve: $\boldsymbol{\delta}\mathbf{A}_0 = quadprog(\boldsymbol{\lambda}_0, \widetilde{y}_0)$;
Set: $\mathbf{s}_0 = (\mathbf{U}\mathbf{A}_0 - \mathbf{A}_{\max}/\|\mathbf{U}\mathbf{A}_0 - \mathbf{A}_{\max}\|)$;
Set: $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_0 + \alpha_0 \cdot \mathbf{s}_0$;
while $|L(\boldsymbol{\lambda}_{k+1}) - L(\boldsymbol{\lambda}_k)| > TOL$
  $\mathbf{s}_k = \mathbf{U}\mathbf{A}_k - \mathbf{A}_{\max}/\|\mathbf{U}\mathbf{A}_k - \mathbf{A}_{\max}\|$;
  $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \alpha_k \cdot \mathbf{s}_k$;
  $\boldsymbol{\delta}\mathbf{A}_k = quadprog(\boldsymbol{\lambda}_k, \widetilde{y}_k)$;
  $\mathbf{A}_{k+1} = \mathbf{A}_k + \boldsymbol{\delta}\mathbf{A}_k$;
  Update $(G_{\rm ap})_{k+1}$ and $(C_{\rm ap})_{k+1}$ with $\mathbf{A}_{k+1}$;
  Solve $\widetilde{y}_{k+1}$ using **(20)** with updated macromodel;
  $k = k + 1$;
end while

However, directly solving the large (4) is still inefficient during the sequential programming. The key to this problem is to efficiently calculate and update the sensitivities $y^{(1)}$ and $y^{(2)}$. As discussed in Section IV, this can be resolved by a structured and parameterized model order reduction. In addition, a detailed outline of this Algorithm will be presented in Section V.

IV. Sensitivity By Structured and Parameterized Macromodel

In this section, we will show that the nominal temperature and its sensitivities can be separately obtained by a structured and parameterized model order reduction, which is general for any linear network. We then apply this technique to obtain a structured and parameterized macromodel for the thermal *RC* network. Here, the parameter to be expanded is the thermal-via density A_i.

A. Structured and Parameterized Model Order Reduction

Assuming the impact by the geometrical-parameter perturbation to the perturbation at output (sensitivity) is much smaller than the one by the frequency perturbation, the temperature state variable $x(A_1, \ldots, A_K, s)$ can be approximated by the Taylor expansion with respect to only geometrical parameters $\mathbf{A}$

$$x(\mathbf{A}, s) = \sum_{i_1}^{\infty} \cdots \sum_{i_K}^{\infty} x_{1,\ldots,K}^{(i_1+\cdots+i_K)}(s)(\delta A_1)^{i_1} \cdots (\delta A_K)^{i_K}. \tag{16}$$

Substituting (16) into (4), and explicitly matching the moment on both sides for each A_i up to the second order, we can reformulate (4) into an augmented and parameterized state-equation

$$(G_{\rm ap} + sC_{\rm ap})x_{\rm ap} = B_{\rm ap}u(t), \quad y_{\rm ap} = L_{\rm ap}^T x_{\rm ap} \tag{17}$$

with

$$G_{\rm ap} = \begin{bmatrix} G_0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ A_1 g_1 & G_0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ A_K g_K & 0 & \ldots & G_0 & 0 & 0 & \ldots & 0 \\ 0 & A_1 g_1 & 0 & \ldots & G_0 & 0 & \ldots & 0 \\ 0 & A_2 g_2 & A_1 g_1 & 0 & \ldots & G_0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A_k g_K & \ldots & 0 & \ldots & G_0 \end{bmatrix} \tag{18}$$

and

$$\begin{aligned} x_{\rm ap} &= \left[x_0^{(0)}, x_1^{(1)}, \ldots, x_K^{(1)}, x_{1,1}^{(2)}, \ldots, x_{K,K}^{(2)}\right]^T \\ B_{\rm ap} &= [B, 0, \ldots, 0, 0, \ldots, 0]^T \\ L_{\rm ap} &= [L, \delta A_1 L, \ldots, \delta A_K L, \delta A_1 \delta A_1 L, \ldots, \delta A_K \delta A_K L]^T. \end{aligned}$$

Note that $C_{\rm ap}$ has the same lower-triangular structure as $G_{\rm ap}$ does.

The system state variable $y_{\rm ap}$ at output for those critical tiles can be also divided into three parts: nominal value $y^{(0)} = y_0^{(0)}$ $(\in R^1)$, first-order sensitivity $y^{(1)} = \left\{y_1^{(1)}, \ldots, y_K^{(1)}\right\}$ $(\in R^K)$, and second-order sensitivity $y^{(2)} = \left\{y_{1,1}^{(2)}, \ldots, y_{K,K}^{(2)}\right\}$ $(\in R^{K\times K})$. As a result, solving (17) results in the nominal value of temperature $y^{(0)}$, and its according to first-order sensitivity $y^{(1)}$ and second-order sensitivity $y^{(2)}$ with respect to each parameter A_i.

Because the dimension of the system (17) is large, its order needs to be reduced by projection. To have a dimension reduced macromodel with preserved moments up to qth order, a small dimensioned projection matrix V can be constructed recursively using the Arnoldi method [26] to (17). As the obtained V is flat and has no structure, directly projecting (17) by V leads to a reduced macromodel losing the lower-triangular block structure of $G_{\rm ap}$ and $C_{\rm ap}$. As a result, $y^{(0)}$, $y^{(1)}$, and $y^{(2)}$ are coupled with each other and cannot be solved separately.

Instead of using the flat projection matrix V, we introduce a structured projection matrix

$$\mathcal{V} = \mathrm{diag}[V_0, \underbrace{V_1, \ldots, V_K}_{K}, \underbrace{V_{K+1}, \ldots, V_{K^2}}_{K^2}] \tag{19}$$

by partitioning V according to the dimension of $x^{(0)}$, $x^{(1)}$, and $x^{(2)}$, and stacking the partitioned blocks into a block-diagonal form, where each V_i $(i = 0, 1, \ldots, K, K+1, \ldots, K^2)$ is further orthonormalized with each other.

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.

As a result, the order-reduced state matrices become

$$\begin{aligned}\widetilde{G}_{\rm ap} &= \mathcal{V}^T G_{\rm ap}\mathcal{V}\\ \widetilde{C}_{\rm ap} &= \mathcal{V}^T C_{\rm ap}\mathcal{V}\\ \widetilde{B}_{\rm ap} &= \mathcal{V}^T B_{\rm ap}\\ \widetilde{L}_{\rm ap} &= \mathcal{V}^T L_{\rm ap}.\end{aligned}$$

The structured and parameterized macromodel

$$\widetilde{H}_{\rm ap} = \widetilde{L}_{\rm ap}(\hat{G}_{\rm ap} + s\hat{C}_{\rm ap})^{-1}\widetilde{B}_{\rm ap}$$

has the following property.

Theorem 1: The first q block moments expanded at s_0 are identical for $\widetilde{H}_{\rm ap}(s)$ and $H(s)$.

Proof: Because $span\{V\} \subseteq span\{\mathcal{V}\}$, a qth ordered projection by $\mathcal{V}$ still preserves at least q moments according to [23].

B. Sensitivity Generation

The time-domain transient response of the reduced model can be solved by Backward–Euler method. The reduced system equation at the time-instant t with time-step h is

$$\begin{aligned}\left(\widetilde{G}_{\rm ap} + \frac{1}{h}\widetilde{C}_{\rm ap}\right)\widetilde{x}_{\rm ap}(t) &= \frac{1}{h}\widetilde{C}_{\rm ap}\widetilde{x}_{\rm ap}(t-h) + \widetilde{B}_{\rm ap}u(t)\\ \widetilde{y}_{\rm ap}(t) &= \widetilde{L}_{\rm ap}^T\widetilde{x}_{\rm ap}(t).\end{aligned} \tag{20}$$

where

$$\widetilde{G}_{\rm ap}=\begin{bmatrix} \widetilde{G}_0 & 0 & \dots & 0 & 0 & 0 & \dots & 0\\ A_1\widetilde{g}_1 & \widetilde{G}_0 & \dots & 0 & 0 & 0 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots\\ A_K g_K & 0 & \dots & \widetilde{G}_0 & 0 & 0 & \dots & 0\\ 0 & A_1\widetilde{g}_1 & 0 & \dots & \widetilde{G}_0 & 0 & \dots & 0\\ 0 & A_2 g_2 & A_1\widetilde{g}_1 & 0 & \dots & \widetilde{G}_0 & \dots & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & A_k\widetilde{g}_K & \dots & 0 & \dots & \widetilde{G}_0 \end{bmatrix} \tag{21}$$

and

$$\begin{aligned}\widetilde{y}_{\rm ap} &= \left[\widetilde{y}^{(0)}, \widetilde{y}^{(1)}, \widetilde{y}^{(2)}\right]^T\\ &= \left[\widetilde{y}_0^{(0)}, \widetilde{y}_1^{(1)}, \dots, \widetilde{y}_K^{(1)}, \widetilde{y}_{1,1}^{(2)}, \dots, \widetilde{y}_{K,K}^{(2)}\right]^T.\end{aligned}$$

Note that the reduced $\widetilde{C}_{\rm ap}$ has the same structure as $\widetilde{G}_{\rm ap}$.

Since the reduction preserves the block structure, the reduced nominal value $\widetilde{y}^{(0)}$, first-order sensitivity $\widetilde{y}^{(1)}$ and second-order sensitivity $\widetilde{y}^{(2)}$ at output (critical tiles) can be solved independently. Then, the temperature profile at those critical tiles perturbed by the parameter is their summation

$$\widetilde{y}(\mathbf{A}, t) = \widetilde{y}^{(0)}(\mathbf{A}, t) + \widetilde{y}^{(1)}(\mathbf{A}, t) + \widetilde{y}^{(2)}(\mathbf{A}, t). \tag{22}$$

The advantages of such a structured and parameterized model order reduction are twofold. First, because the projection (19) preserves the lower-triangular structure, (20) can be efficiently solved using the block-back-substitution. There is only one factorization cost from the diagonal block, i.e., the reduced block of nominal state matrix

$$\widetilde{G}_0 + \frac{1}{h}\widetilde{C}_0.$$

Second, the separately solved sensitivity can be utilized during any gradient-based optimization procedure including the sequential programming as discussed in Section III.

V. Algorithm and Experiments

A. Overall Algorithm

The procedure of the sequential subgradient optimization is outlined in Algorithm 1. Note that α_k is the step size usually determined through a geometric regression procedure [28].

The structured and parameterized macromodel provides a convenient interface between the simulation and the optimization. This can be understood from two aspects. Firstly, the reduction and the solving of the nominal response are both one-time computation in practice. The reduced state matrices can be repeatedly used, and only the sensitivity needs to be solved and updated (for a new $\mathbf{A}$) during each iteration. Next, since the reduced model is much smaller than the original one, its nominal value and sensitivities can be efficiently solved from (20). As shown by experiments, the optimization procedure in Algorithm 1 is computationally efficient compared to using the direct matrix-solver.

B. Numerical Results

Our structured and parametrized macromodeling (called SP-Macro) and thermal-via allocation are implemented in MATLAB and C++, and run on Linux workstation with Intel Pentium IV 2.66G CPU and 2G RAM. The examples have the following settings. k_1 (thermal conductive constant) is 100 W/m·K for silicon and 400 W/m·K for copper, and k_2 (thermal capacitive constant) is 1.75×10^6 J/m$^3\cdot$K for silicon, and 3.55×10^6 J/m$^3\cdot$K for copper. The substrate is 500 μm thick, the device layer is 6 μm thick, and interlayer thickness is 1 μm thick. Four silicon layers are used and the thermal-via is assumed to be copper. The unit via area is 2×2 μm^2. The overall chip size is 2×2 cm^2, and the number of individual modules and its according size are from MCNC benchmarks. We increase the model complexity by increasing the number of discretized tiles and the number of critical tiles. The critical titles are selected manually according to the functionary/reliability of benchmark circuit and hence may show a differently increasing rate.

The power distribution at each title is chosen similarly as [16], where 90% of tiles have power densities from 0 to 2×10^6 W/m^2, and their clock gating pattern has a period of 500 ms, where the power in the standby mode is 5% of the running mode. The other 10% of tiles having power densities from 3×10^6 W/m^2 to 9×10^6 W/m^2, and their clock gating pattern has a period of 250 ms, where the power in the standby mode is 20% of the running mode. In addition, note that a

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.

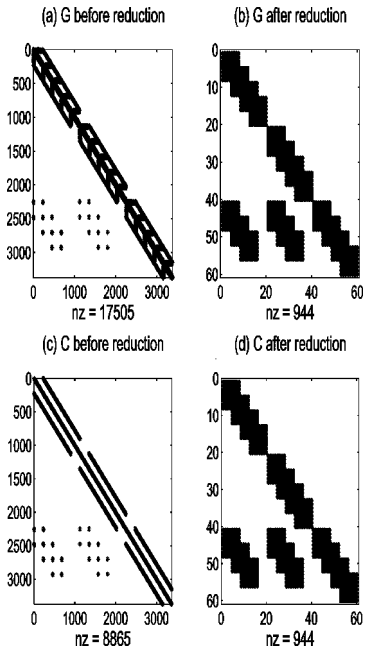

Fig. 5. Nonzero (NZ) pattern of the parameterized state matrices G and C before and after the structured model order reduction.

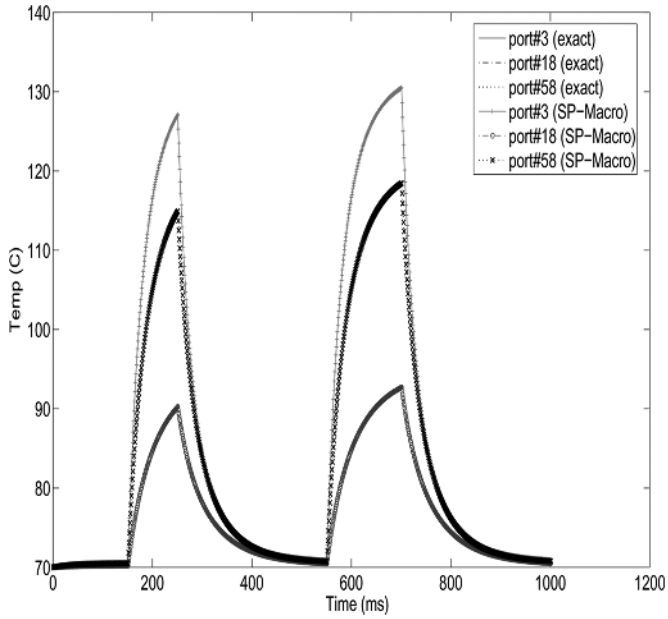

Fig. 6. Transient temperature responses of exact and structured and parameterized macro (SP-Macro) models at port 3, 18, and 58 of layer-1 with step-response input. The macromodels are visually identical to those exact models.

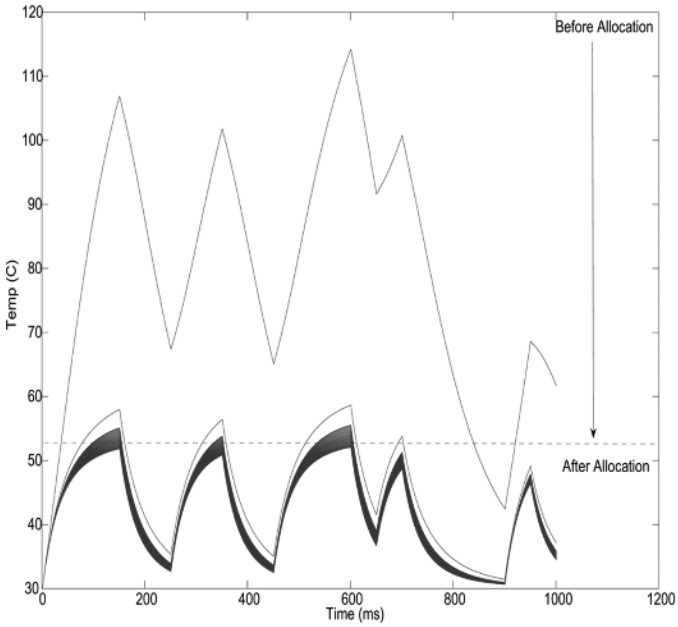

Fig. 7. Iterative optimizations showing the hotspot reduction by thermal-via allocation under the input of transient thermal-power signature at port 32 of layer-1. The ceiling temperature is 52 °C.

single-input-multi-output (SIMO) [24] is assumed when the port number in B is large.

1) Structured and Parameterized Macromodel: One detailed 3-D thermal *RC* circuit is used to verify the proposed algorithm. It has four layers and each layer contains about 1 K tiles. 64 tiles of each layer are selected as critical tiles. The total thermal-via density constraint is 3000, and the local via number constraint is randomly generated from 50 to 400. Structured and parameterized model reduction is first applied to generate SP-Macro for the thermal-via allocation considering the transient effect. Then, the entire circuit is used to generate the steady-state map of the temperature profile.

For the SP-Macro and original models, Fig. 5 shows the parameterized state-matrix structure before and after the reduction. The parameterized state matrix shows a lower-block triangular structure, and the structured reduction preserves such a low-block triangular structure. As a result, the reduced model can be solved efficiently by the backward substitution with only one factorization cost, coming from the reduced nominal state matrix in the diagonal. As shown below, it is efficient to apply such a structured and parameterized macromodel for the sake of iterative optimizations.

Moreover, Fig. 6 compares the time-domain transient temperature at selected three critical tiles (3, 18, 58) using (22). Sixteen moments are used for the moment matching. Clearly, the reduced models are visually identical to original ones.

2) Sequential Programming: Furthermore, for the same 3-D thermal circuit above, Fig. 7 shows the successive temperature cooling by allocating the thermal-via according to the calculated transient sensitivity. The thermal-violation integral is minimized until the ceiling temperature is 52 °C is met. In addition, Fig. 8 shows the subgradient optimization procedure after four iterations, where the dual problem quickly converges with the primal problem at one normalized value 0.7. Clearly, our sequential programming could effectively minimize the thermal-violation integral.

3) Thermal-Via Allocation: We then study the effectiveness of our thermal-via allocation algorithm by comparing it with the one using the steady-state analysis. Note that directly solving steady-state equation cannot generate the sensitivity for the optimization. Therefore, during the allocation with use of the steady-state analysis, the parameterized state (17) in the steady-state is used to calculate the steady-state response and

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.

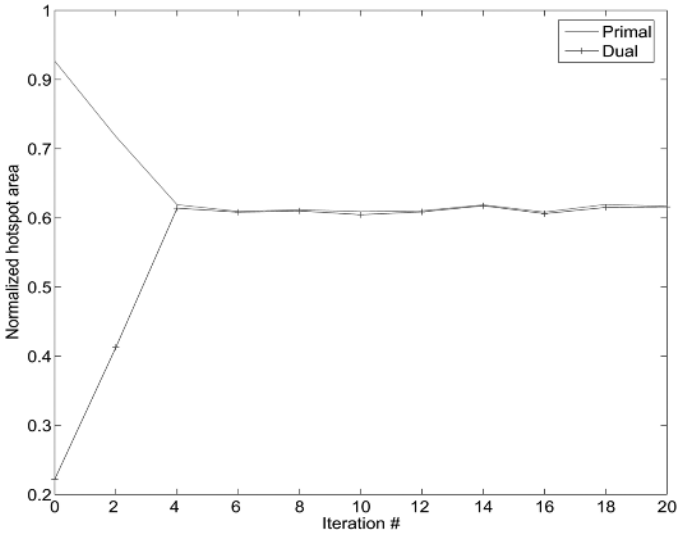

Fig. 8. Convergence of subgradient optimization of primal and dual problems. The hotspot is represented by violation integral normalized to the maximum. α_0 here is set to 0.7.

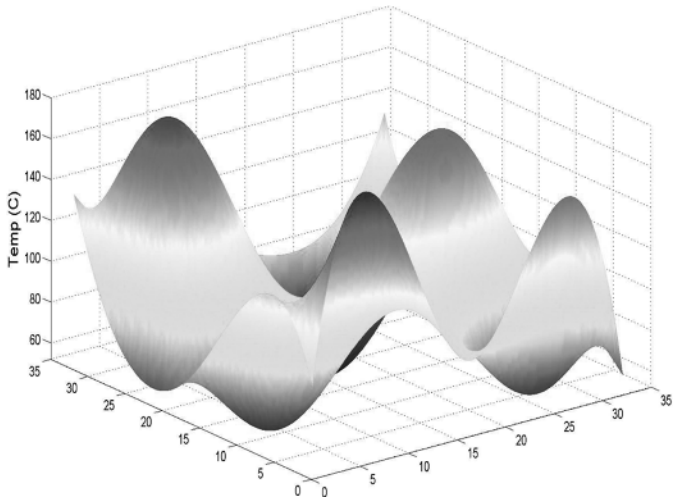

Fig. 10. Steady-state temperature map of top layer (layer-1) before thermal-via allocation.

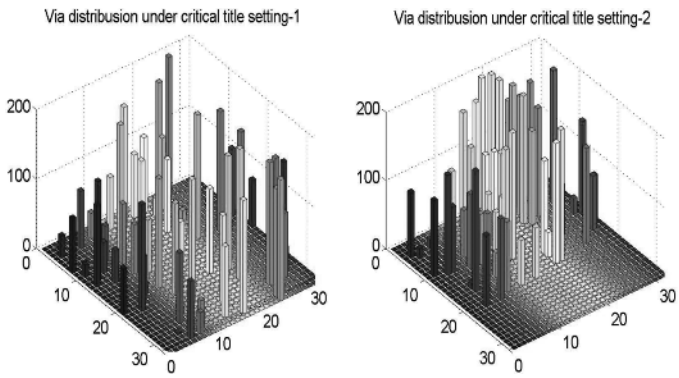

Fig. 9. Thermal-via distribution under two different settings of critical titles.

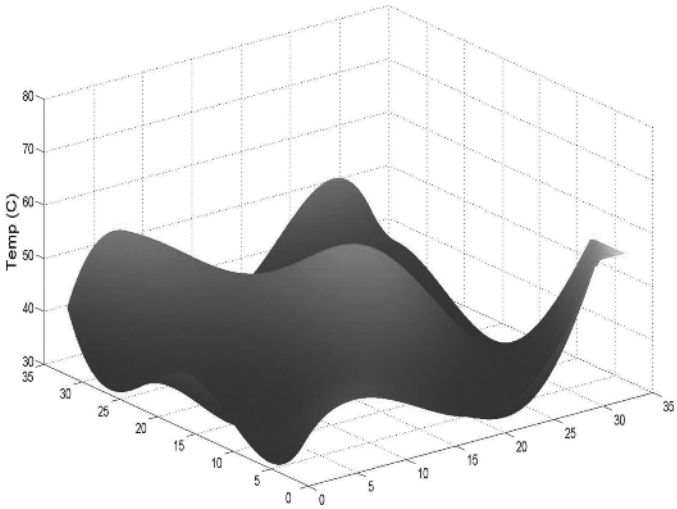

Fig. 11. Steady-state temperature map of top layer (layer-1) after thermal-via allocation using transient temperature profile.

the sensitivity. Then, as discussed in Section III-A (see Fig. 4), a rectangular area is used as the objective, formed by the "steady" maximum temperature and the ceiling temperature. In contrast, our proposed method with use of the transient analysis utilizes a "dynamic" violation integral. Finally, both of the steady-state and transient formulations are solved by the sequential programming.

For the same 3-D thermal circuit above, Fig. 9 further shows the allocated via density distribution. There are two power inputs injected with different clock-gating-periods to study the impact of critical tiles. It results in a different set of locations for those critical tiles. As a result, it leads to two different distributions of allocated vias in Fig. 9. Therefore, the worst-case dynamic thermal-power-input or workload needs to be assumed to the worst-case temperature, which accordingly determines the critical tiles.

Moreover, Figs. 10 and 11 further show the steady-state temperature map across the top layer (layer-1). The initial chip temperature at the top layer is around 150 °C, and its temperature profile at steady-state is shown in Fig. 10. In contrast, the allocation results in a cooled temperature profile that closely approaches the ceiling temperature as shown in Fig. 11. In addition, note that because the transient thermal-violation integral is used as the figure of merit, the spatial distribution of allocated thermal-via shows a large difference from the temperature hot-spots at the steady state.

Table II further analyzes the runtime scalability and allocated thermal-via density by the proposed method and the steady-state analysis. Because directly solving steady-state equation needs to handle large sized matrix, it consumes runtime and memory during the sequential optimization. In contrast, the macromodel can efficiently match the transient response using around 20 moments. For a circuit with 8192 tiles, our model reduces runtime by 126× (62's versus 7809's) compared to the steady-state analysis. More importantly, due to the use of an accurate figure of merit, the transient thermal-violation integral, our allocated thermal-via density is much smaller than the one by steady-state

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.

TABLE II
EXPERIMENT SETTING AND RESULTS OF THERMAL-VIA PLANNING TIME AND NUMBER. THE ALLOCATED THERMAL-VIA OF STEADY-STATE ANALYSIS IS BASED ON THE REDUCED MACROMODEL WITH THE USE OF THERMAL-VIOLATION INTEGRAL DEFINED BY THE MAXIMUM TEMPERATURE

total/critical tile#	global via bound	original/ceiling T ($^\circ C$)	Steady-state(direct)				Transient(SP-macro)				
			solve dc (s)	solve tran (s)	allo-via	opt-T($^\circ C$)	redu ckt (s)	solve sens (s)	qp-prog plan (s)	allo-via	opt-T($^\circ C$)
256/30	704	120/40	1.64	10.27	440	40.4	0.12	0.19	0.15	360	40.2
1024/60	2818	120/40	12.62	130.12	2281	41.5	1.08	0.96	0.42	1609	41.7
4096/80	5980	140/50	341.13	3872.98	5620	52.1	12.92	6.28	1.92	3217	51.9
8192/100	8218	140/50	7809.12	NA	8021	53.3	46.27	16.92	8.98	4382	53.1
16384/120	18000	160/60	NA	NA	17600	63.6	120.89	101.23	23.65	9280	63.4
32768/200	24000	160/60	NA	NA	23800	65.4	262.12	257.21	42.78	11660	65.3

analysis under the same targeted ceiling temperature. For a circuit with 32 768 tiles, our design reduces 2.04× (11 660 versus 23 800) thermal vias compared to the steady-state analysis.

VI. CONCLUSION AND DISCUSSIONS

The previous thermal-via allocations [16], [17] for 3-D IC employ the direct steady-state analysis, ignore the temporal and spatial variations of the thermal-power, and hence, may result in the excessive number of thermal vias. In this paper, to consider the temporally and spatially variant thermal-power, a thermal-violation integral of the transient temperature is proposed to accurately capture the thermal violation, and a nonlinear optimization is then used to minimize the thermal-violation integral. The nonlinear programming can be solved through the sequential quadratic programing, where sensitivities are calculated and updated efficiently from a structured and parameterized macromodel. Experiments show that compared to the existing method using the steady-state thermal analysis, our method is 126 faster to obtain the temperature profile, and reduces the number of thermal vias by 2.04 under the same temperature bound.

Clearly, the proposed structured and parameterized macromodel can be used for a number of integrity-driven physical-design problems. For example, we have recently presented a 3-D via-planning for simultaneous power and thermal integrity [29], where vias are allocated to satisfy constraints on power resonance of power/ground planes in the package and constraints on maximum temperature in stacked IC dices. Again, the structured and parameterized macromodel is utilized to develop an efficient yet effective algorithm, which reduces via number compared to the sequential power and thermal integrity optimization.

Note that a "dynamic" thermal-violation integral for the thermal-integrity is used in this paper instead of using a "steady" maximum temperature. Both the "dynamic" thermal-violation integral and the "steady" maximum temperature can be obtained from the worst-case temperature profile. As discussed in [19], when the workload is available the worst-case temperature profile and its associated critical titles can be both characterized from the thermal-power, and the guard-land can be used to avoid the under-design. However, it is computationally expensive if not possible to determine the worst-case temperature profile from all kinds of dynamic workloads. The stochastic characterization approach such as the principal component analysis can be applied to find a set of principal temperature and its associated critical tiles.

In addition, we assume that the thermal vias are aligned for all layers in this paper. Though the proposed approach is general to consider the nonaligned vias, it may introduce additional cost to build the parameterized macromodel to provide more design freedoms. In the future, we will study a layer-wised via-relocation to incrementally transform the parameterized model with aligned vias into the one with nonaligned vias by perturbation.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their insightful comments to make this paper better.

REFERENCES

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

[2] S. Das, "Design automation and analysis of three dimensional integrated circuits," Ph.D. dissertation, EECS Dept., Massachusetts Inst. Technol., Boston, 2004.

[3] W. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A.M. Sule, M. Steer, and P.D. Franzon, "Demystifying 3D ICs: the pros and cons of going vertical," *IEEE Des. Test Comput.*, vol. 22, no. , pp. 498–510, Dec. 2005.

[4] S. Lim, "Physical design for 3D system-on-package: Challenges and opportunities," *IEEE Des. Test Comput.*, vol. 22, no. 6, pp. 532–539, Dec. 2005.

[5] C. C. Teng, Y. K. Cheng, E. Rosenbaum, and S. M. Kang, "ITEM: A temperature-dependent electromigration reliability diagnosis tool," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 16, no. 8, pp. 882–893, Aug. 1997.

[6] K. Banerjee, A. Mehrotra, A. Sangiovanni-Vincentelli, and C. Hu, "On thermal effects in deep sub-micron VLSI interconnects," in *Proc. ACM/IEEE Des. Autom. Conf. (DAC)*, 1999, pp. 885–891.

[7] T. Wang and C. Chen, "Thermal-adi: A linear-time chip-level dynamic thermal simulation algorithm based on alternating-direction-implicit (ADI) method," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 11, no. 4, pp. 691–700, Aug. 2003.

[8] M. R. Stan, K. Skadron, M. Barcella, W. Huang, K. Sankaranarayanan, and S. Velusamy, "Hotspot: a dynamic compact thermal model at the processor-architecture level," *Microelectron. J.*, vol. 34, pp. 1153–1165, 2003.

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

[10] P. Liu, Z. Qi, H. Li, L. Jin, W. Wu, S. X.-D. Tan, and J. Yang, "Fast thermal simulation for architecture level dynamic thermal management," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2005, pp. 639–644.

[11] S. Lin and K. Banerjee, "An electrothermally-aware full-chip substrate temperature gradient evaluation methodology for leakage dominant technologies with implications for power estimation and hot-spot management," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2006, pp. 568–574.

[12] Y. Yang, C. Zhu, Z. Gu, L. Shang, and R. P. Dick, "Adaptive multi-domain thermal modeling and analysis for integrated circuit synthesis and design," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2006, pp. 575–582.

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.

[13] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. Stan, "Hotspot: a compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 15, no. 5, pp. 501–513, May 2006.
[14] 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. (ICCAD)*, 2003, pp. 86–89.
[15] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2004, pp. 306–313.
[16] B. Goplen and S. Sapatnekar, "Thermal via placement in 3D ICs," in *Proc. Int. Symp. Phys. Des. (ISPD)*, 2005, pp. 167–174.
[17] J. Cong and Y. Zhang, "Thermal via planning for 3D ICs," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2005, pp. 745–752.
[18] Z. Li, X. Hong, Q. Zhou, H. Yang, V. Pitchumani, and C. Cheng, "Integrating dynamic thermal via planning with 3D floorplanning algorithm," in *Proc. Int. Symp. Phys. Des. (ISPD)*, 2006, pp. 178–185.
[19] V. Tiwari, D. Singh, S. Rajgopal, G. Mehta, R. Patel, and F. Baez, "Reducing power in high-performance microprocessors," in *Proc. ACM/IEEE Des. Autom. Conf. (DAC)*, 1998, pp. 732–737.
[20] K. Skadron *et al.*, "Temperature-aware microarchitecture," in *Proc. Int. Symp. Comput. Arch.*, 2003, pp. 2–13.
[21] W. Liao, L. He, and K. Lepak, "Temperature and supply voltage aware performance and power modeling at microarchitecture level," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 24, no. 7, pp. 1042–1053, Jul. 2005.
[22] H. Yu, Y. Shi, L. He, and T. Karnik, "Thermal via allocation for 3D ICs considering temporally and spatially variant thermal power," in *Proc. Int. Symp. Low Power Electron. Des. (ISLPED)*, 2006, pp. 156–161.
[23] E. J. Grimme, "Krylov projection methods for model reduction," Ph.D. dissertation, Math. Dept., Univ. Illinois Urbana-Champaign, Urbana-Champaign, 1997.
[24] A. Odabasioglu, M. Celik, and L. Pileggi, "Prima: Passive reduced-order interconnect macro-modeling algorithm," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 17, no. 8, pp. 645–654, Aug. 1998.
[25] L. Daniel, O. C. Siong, L. S. Chay, K. H. Lee, and J. White, "A multiparameter moment matching model reduction approach for generating geometrically parameterized interconnect performance models," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 23, no. 5, pp. 678–693, May 2004.
[26] X. Li, P. Li, and L. Pileggi, "Parameterized interconnect order reduction with explicit-and-implicit multi-parameter moment matching for inter/intra-die variations," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2005, pp. 806–812.
[27] C. Visweswariah, R. A. Haring, and A. R. Conn, "Noise considerations in circuit optimization," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 19, no. 6, pp. 679–690, Jun. 2000.
[28] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*. New York: Wiley, 1993.
[29] H. Yu, J. Ho, and L. He, "Simultaneous power and thermal integrity driven via stapling in 3D ICs," in *Proc. Int. Conf. Comput. Aided Des. (ICCAD)*, 2006, pp. 802–808.

Hao Yu (S'02–M'06) received the Ph.D. degree in electrical engineering from University of California, Los Angeles (UCLA), in 2006.

He was a Graduate Student Researcher with the Electrical Engineering Department, UCLA, between 2002 and 2006, and has been a Senior Member of the technical staff developing the fastest analog/RF simulation at Berkeley Design Automation since 2006. His research interests include numerical prototyping of high-performance mixed-mode circuits: structured and parameterized model order reduction and fast differential/integral equation solvers for SPICE and TCAD; and Macromodel-based synthesis platform: integrity and robustness driven hybrid-system-integration from device to system level.

Dr. Yu's work on structured and parameterized macromodeling was nominated as the Best Paper Award in Design Automation Conference (2006) and International Conference of Computer-Aided Design (2006), respectively.

Yiyu Shi received the B.E. degree (with honors) in electronic engineering from Tsinghua University, Beijing, China, in 2005, and the M.S. degree (with honors) in electrical engineering in the University of California, Los Angeles, in 2007, where he is currently pursuing the Ph.D. degree in computer-aided design.

His current research interests include design automation for VLSI circuits and systems and large-scale optimization.

Lei He (S'94–M'99–SM'08) received the Ph.D. degree in computer science from the University of California, Los Angeles (UCLA), in 1999.

He is an Associate Professor with the Electrical Engineering Department, UCLA, and was a faculty member with the University of Wisconsin, Madison, between 1999 and 2001. He also held visiting or consulting positions with Intel, Hewlett-Package, Cadence, Synopsys, Rio Design Automation, and Apache Design Solutions. His research interests include VLSI circuits and systems, and electronic design automation. He has published over 170 technical papers and has been a technical program committee member for a number of conferences including Design Automation Conference, International Conference on Computer-Aided Design, International Symposium on Low Power Electronics and Design, and International Symposium on Field Programmable Gate Array.

Dr. He was a recipient of the National Science Foundation CAREER Award in 2000, the UCLA Chancellor's faculty career development award (highest class) in 2003, the IBM Faculty Award in 2003, the Northrop Grumman Excellence in Teaching Award in 2005, the Best Paper Award at the 2006 International Symposium on Physical Design, and multiple Best Paper Nominations at Design Automation Conference and International Conference on Computer-Aided Design.

Tanay Karnik (M'88–SM'04) received the Ph.D. degree in computer engineering from the University of Illinois at Urbana-Champaign, Urbana-Champaign, in 1995. From 1995 to 1999, he worked with the Strategic CAD Lab, Intel. Since March 1999, he has lead the power delivery, soft error rate, and low power circuits research in the Circuits Research Lab, Intel. His research interests include the areas of variation tolerance, power delivery, soft errors, and physical design. He has published over 40 technical papers, has 36 issued and 40 pending patents in these areas.

Dr. Karnik was a recipient of an Intel Achievement Award for the pioneering work on integrated power delivery. He has presented several invited talks and tutorials, and has served on 5 Ph.D. students' committees. He was a member of DAC, ICCAD, ICICDT, and ISQED program committees and JSSC, TCAD, TVLSI, TCAS review committees. He was a General Chair of ISQED'08 and ICICDT'08.

Authorized licensed use limited to: Nanyang Technological University. Downloaded on March 07,2010 at 22:45:58 EST from IEEE Xplore. Restrictions apply.