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

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

Abstract— **The existing 3D 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 using macromodel by 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 using sensitivities from the macromodel. Compared to the existing method using the steady-state thermal analysis, our method in experiments is 126X faster to obtain temperature, and reduces the number of thermal vias by 2.04X under the same temperature bound.**

Index Terms— **Cooling technology, Thermal power management, 3D IC design, Structured and parameterized macromodel, Sequential programming**

I. INTRODUCTION

The existing two-dimensional (2D) high-performance system-on-chip (SoC) design is limited by the interconnect delay and device density. Three-dimensional (3D) 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 3D IC design including the active device layers, vertical through vias, and the substrate. Due to the increased power density, heat removal is extremely important in 3D-ICs [1]. It is well known that excessively high temperature can significantly degrade interconnect and device reliability, and cause functional or timing failures through the electro-thermal 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 3D 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 3D thermal via planning is thereby to allocate vias in order to 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 the running average of the cycle-accurate (nanosecond) power 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], [17] have 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], [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 RC model to calculate the transient temperature. However, [8], [13], [16], [17] directly solve the matrix-formed state equation. It is inefficient to calculate the nominal temperature and its sensitivity with respect to the thermal-via density for large scale designs. Moreover, there is no effective interaction between the simulator and the optimizer. The design procedure is either based on iterations [16], or based on an approximated square-root relation [17] between temperature and thermal-vias. It may not converge or may lead to inaccurate results. As a result, it is needed to develop an accurate and efficient solution to consider the transient temperature and to help the optimization with use of 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 using 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 sensitives provided by the macromodel. Experiments show that compared to the steady-state thermal analysis, our method is 126X faster to obtain the

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

Yiyu Shi and Lei He are with Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: {yshi,lhe}@ee.ucla.edu).

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

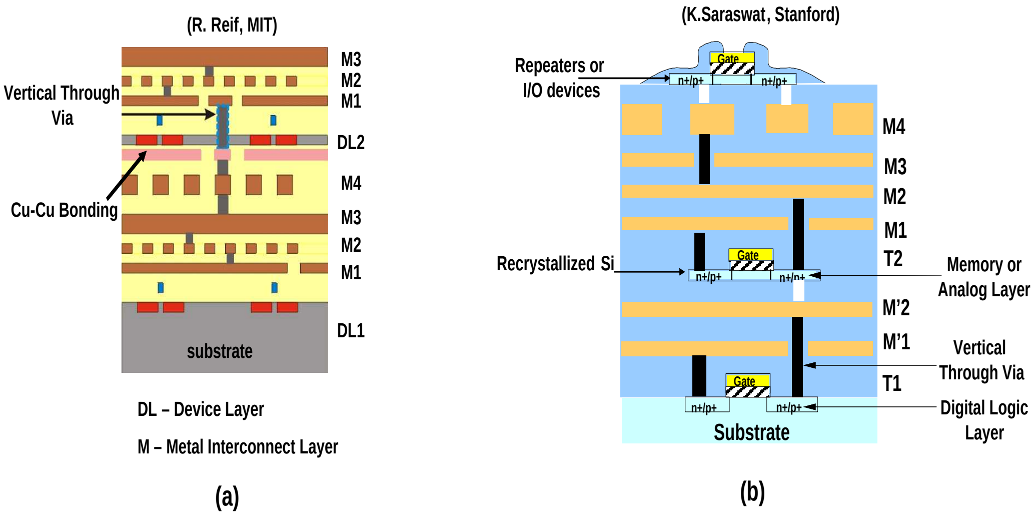

Fig. 1. The typical 3D IC techniques including: (a) Cu-Cu wafer bonding and (b) crystallization of α-Si. They both have active device layers, inter-layer dielectrics, vertical through vias, and the substrate.

temperature profile, and reduces the number of thermal vias by 2.04X under the same temperature bound.

The rest of the paper is organized as follows. In Section II, we present the preliminary for 3D thermal model and analysis. In Section III, we formulate a nonlinear optimization to accurately allocate 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 efficiently 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 was presented in [22].

II. 3D Thermal Model and Analysis

In this Section, we present how to build a parameterized thermal model for 3D 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 though a pair of thermal resistance and capacitance driven by the current source, modeling 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 1D resistor network with attached external voltage sources to model the ambient temperature.

Each active device layer and the inter-layer dielectric in 3D layouts can be uniformly discretized into N tiles by the finite difference method. As shown in Fig. 2, in steady-state analysis, tiles connected by thermal resistance R. Heat sources modeled as time-invariant current sources. Steady-state temperature can be obtained by directly solving a time-invariant linear equation. In contrast, as for the transient analysis, tiles connected by thermal resistance and capacitance RC. Heat sources modeled as time-variant current sources. Usually, the granularity of discretized 3D IC smaller than thermal space constants might not be necessary [13]. Because the proposed method is targeted at 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 3D IC usually requires to consider many heterogeneous components in one system, it can lead to a more complicated thermal model than that for the 2D ICs.

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)

In addition, there are two types of specified tiles: *critical tiles* and *input tiles*. Critical tiles are those tiles with hottest temperatures that can cause thermal violations leading to the 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 heat-dissipation $u(t)$ 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

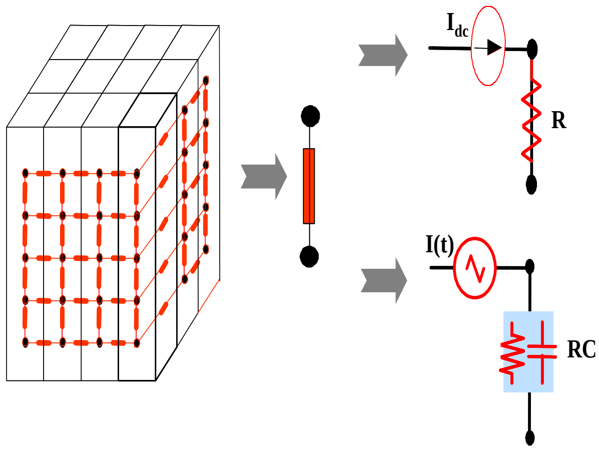

Fig. 2. The thermal models extracted for the steady-state analysis and the transient analysis.

the thermal-via density ρ_i by

$$\rho_i = A_i/a,$$

where a is the unit area of thermal-via determined by the process. 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 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 topological matrix g_i/c_i for one inserted unit-via conductance/capacitance

$$g_i = (k_1/t)X_i \qquad c_i = (k_2 t)X_i,$$

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

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

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. Moreover, $\sum_{i=1}^{K} A_i g_i$ 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.

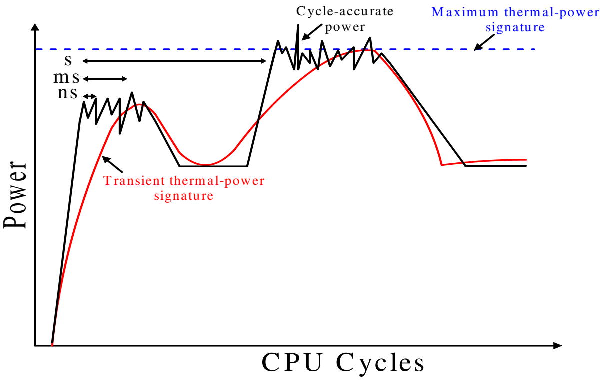

Fig. 3. The definitions of the cycle-accurate power, transient thermal-power signature, and maximum thermal-power signature at the different scale of time constant.

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

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

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

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

where $\mathbf{A} = [A_1, ..., A_K]$ 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 thermal-power input. As defined in [19] (See Fig. 3), a *transient thermal power* is the running average of the cycle-accurate (often in the range of ns) power over several thermal time constants (often in the range of ms), 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, the (thermal) power density is *time-variant* [8], [19]–[21]. As a result, the temperature varies not only spatially but also temporally. In addition, the temporal variation of temperature can also result from different applications (workloads). Therefore, 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. However, it is rare 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 motivation of this paper to study the via planning problem using a dynamic or transient thermal model.

B. Macromodel by Moment Matching

On the the hand, blindly applying the thermal transient analysis is expensive because it is not efficient to solve (4) for large sized N thermal circuits. Similar to the macromodeling for the interconnect network, the moment matching based model order reduction can be used to obtain a 3D IC macromodel with compact sized q ($q << N$), which not only has a smaller matrix size but also preserves the dominant system response.

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

$$\mathcal{A} = (G + s_0C)^{-1}C \quad \mathcal{R} = (G + s_0C)^{-1}B,$$

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

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

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

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

where $n = \lfloor q/p \rfloor$. By applying the Block Arnoldi orthonormalization [24], the spanned subspace by a smalled dimensioned projection matrix V ($\in R^{N \times q}$) can be found to contain the Krylov subspace

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

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

$$\hat{G} = V^TGV, \; \hat{C} = V^TCV, \; \hat{B} = V^TB, \; \hat{L} = V^TL$$

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 one selected frequency 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 moment matching [23], [24]. The error of reduced model depends on the selection of reduced order q with a detailed analysis of numerical error bound of moment matching in [23].

To further obtain the sensitivity information, the parametrized moments [25] can be obtained by expanding (4) at selected parameter points. However, because the 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. 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 apply those sensitivities in the optimization. This will be addressed in Section IV, and we summarize notations used in this paper below.

- 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)
- $x^{(1)}/y^{(1)}$: $1st$-order sensitivity (at output)
- $x^{(2)}/y^{(2)}$: $2nd$-order sensitivity (at output)
- f: the thermal-violation integral
- $T_{ceiling}$: the targeted ceiling temperature

III. Thermal-Via Allocation Problem

In this Section, to consider an accurate figure of merit for the transient thermal integrity, a thermal-violation integral is first defined, and 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 $T_{ceiling}$:

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

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

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

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 $T_{celling}$. This definition captures the fact that a thermal violation occurs only when the temperature is above the temperature bound for a long enough period. A similar merit is used for noise estimation in [27].

Moreover, the figure of merit for a group of K critical tiles needs 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 addtion to those *local* thermal violation integral f_is. Accordingly, a *thermal violation integral vector* is defined by

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

The thermal-violation integral vector $\mathbf{f}(\mathbf{A})$ is used as an accurate objective function in the sequel to be minimized by allocating thermal vias. Moreover, to compute $f(\mathbf{A})$, the evaluation period t is discretized into finite intervals and Problem 1 becomes semi-definite [27], which can be solved with the provable convergence.

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 over-estimated 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, an nonlinear optimization problem is formulated as

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

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

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

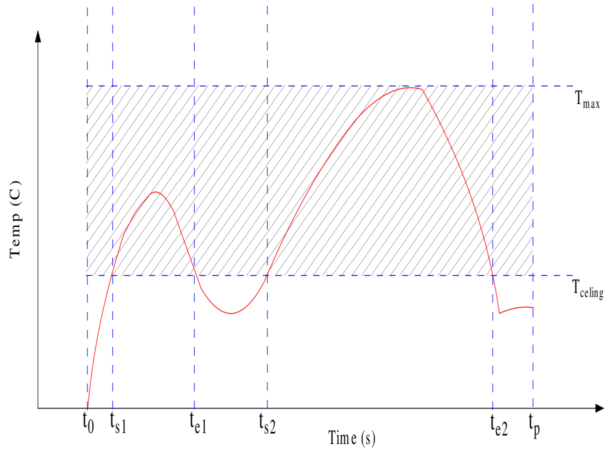

Fig. 4. Figure of merit using thermal-violation integral with defined ceiling temperature under an input of transient thermal-power signature.

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

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

As a result, we have

$$\mathbf{U}\mathbf{A} \le \mathbf{A}_{max}, \tag{11}$$

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

To efficiently solve Problem 1, the below Lagrangian relaxation can be used. The constraint function can be added to the objective function using a vector of Lagrangian multiplier $\lambda = [\lambda_1, ..., \lambda_K]$. As a result, the primal problem (Problem 1) has a following dual problem:

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

where

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

It can be transformed into a sequence of subgradient optimization [28] problems as discussed below.

C. Sequential Programming

In general, $f_k(\mathbf{A})$ is a nonlinear function with respect to all A_is $(i, k = 1, ..., K)$. However, if the via density (area) A_i is only changed by a small amount δA_i $(\delta A_i = A_i - A_i^{(0)})$ 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:

$\underline{Problem\ 3}$:

$$min \quad \nabla \mathbf{f}(\mathbf{A})^T \delta \mathbf{A} + \frac{1}{2}\delta \mathbf{A}^T H \delta \mathbf{A} + \lambda \cdot \nabla \mathbf{h}(\mathbf{A})\delta \mathbf{A}. \tag{15}$$

Note that

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

is the first-order sensitivity, and

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

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

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 called as *sequential quadratic programming* (SQP) [28]. Note that the convergence of the SQP depends on the range of the calculated $\delta \mathbf{A}$. The 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 with the required accuracy.

However, directly solving (4) is still inefficient for such a sequential programming. The key to this problem is to efficiently calculate and update the sensitivities $y^{(1)}$ and $y^{(2)}$. This can be solved by a structured and parameterized model order reduction as discussed below. The 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 separated nominal temperature and its sensitivities can be obtained by a structured and parameterized reduction, which is general for any linear network. We 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. Parameterized and Structured Model Order Reduction

Because the output sensitivity is large with respect to the frequency but small with respect to the geometric parameter, the temperature state variable $x(A_1, ..., A_K, s)$ can be approximated by the Taylor expansion with only respect to the geometrical parameters $\mathbf{A}$:

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

This is similar to the method in [26] modeling process variations for the electrical system. Substituting (16) in (4), and explicitly matching the moment for each A_i up to the second-order, we can reformulate (4) into an augmented and parameterized state equation:

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

with

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

and

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

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

In addition, the system state variable y_{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 $y^{(1)} = \{y_1^{(1)}, ..., y_K^{(1)}\}$ $(\in R^K)$, and second-order sensitivity $y^{(2)} = \{y_{1,1}^{(2)}, ..., y_{K,K}^{(2)}\}$ $(\in R^{K \times K})$. As a result, solving (17) results in the nominal value of temperature $y^{(0)}$, and its according first-order sensitivity $y^{(1)}$ and second-order sensitivity $y^{(2)}$ with respect to each parameter A_i.

Because the dimension of the system equation (17) is large, its order needs to be reduced using projection with preserved moments (of s) up to q-th order. A small dimensioned projection matrix V can be constructed recursively using the Arnoldi method [26]. However, the obtained V has no structure. Directly projecting (17) by V leads to a reduced macromodel losing the lower-triangular block structure of G_{ap} and C_{ap}. In addition, $y^{(0)}$, $y^{(1)}$ and $y^{(2)}$ are coupled with each other and can not be solved separately.

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

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

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. As a result, the order-reduced state matrices become

$$\widetilde{G}_{ap} = \mathcal{V}^T G_{ap} \mathcal{V},\ \widetilde{C}_{ap} = \mathcal{V}^T C_{ap} \mathcal{V},$$
$$\widetilde{B}_{ap} = \mathcal{V}^T B_{ap},\ \widetilde{L}_{ap} = \mathcal{V}^T L_{ap}.$$

In addition, the structured and parameterized macromodel

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

has the following property:

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

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

B. Sensitivity Generation

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

$$(\widetilde{G}_{ap} + \frac{1}{h}\widetilde{C}_{ap})\widetilde{x}_{ap}(t) = \frac{1}{h}\widetilde{C}_{ap}\widetilde{x}_{ap}(t-h) + \widetilde{B}_{ap}u(t)$$
$$\widetilde{y}_{ap}(t) = \widetilde{L}_{ap}^T \widetilde{x}_{ap}(t). \quad (20)$$

where

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

and

$$\widetilde{y}_{ap} = [\widetilde{y}^{(0)}, \widetilde{y}^{(1)}, \widetilde{y}^{(2)}]^T = [\widetilde{y}_0^{(0)}, \widetilde{y}_1^{(1)}, ..., \widetilde{y}_K^{(1)}, \widetilde{y}_{1,1}^{(2)}, ..., \widetilde{y}_{K,K}^{(2)}]^T.$$

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

Because 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. The temperature profile at those critical tiles perturbed by the parameter is

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

The advantages of such a structured and parameterized model order reduction are two fold. Firstly, the nominal response only requires one-time transient simulation, and hence we only need to solve the perturbed response, i.e., the sensitivity, during each iteration. Next, the solved sensitivity can be utilized during any gradient-based optimization procedure including the sequential programming discussed in Section III.

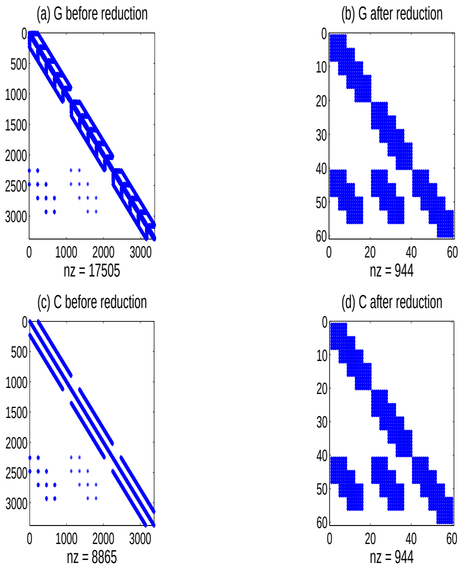

Fig. 5. The non-zero (NZ) pattern of the parameterized state matrices G and C before and after the structured model order reduction.

Algorithm 1 Subgradient Optimization using Structured Parameterized Macromodel

```
Initialize: $(\mathbf{A}_0, \alpha_0, \lambda_0, H_0, k)$;
Solve: $\widetilde{y}_0$ using (20);
Solve: $\delta\mathbf{A}_0 = quadprog(\lambda_0, \widetilde{y}_0)$;
Set: $\mathbf{s}_0 = \frac{\mathbf{UA}_0 - \mathbf{A}_{max}}{||\mathbf{UA}_0 - \mathbf{A}_{max}||}$;
Set: $\lambda_1 = \lambda_0 + \alpha_0 \cdot \mathbf{s}_0$;
while $|L(\lambda_{k+1}) - L(\lambda_k)| > TOL$ do
    $\mathbf{s}_k = \frac{\mathbf{UA}_k - \mathbf{A}_{max}}{||\mathbf{UA}_k - \mathbf{A}_{max}||}$;
    $\lambda_{k+1} = \lambda_k + \alpha_k \cdot \mathbf{s}_k$;
    $\delta\mathbf{A}_k = quadprog(\lambda_k, \widetilde{y}_k)$;
    $\mathbf{A}_{k+1} = \mathbf{A}_k + \delta\mathbf{A}_k$;
    Update $(G_{ap})_{k+1}$ and $(C_{ap})_{k+1}$ with $\mathbf{A}_{k+1}$;
    Solve $\widetilde{y}_{k+1}$ using (20) with updated macromodel;
    $k = k + 1$;
end while
```

V. Algorithm and Experiments

A. Overall Algorithm

The sequential subgradient optimization procedure is outlined in Algorithm 1, where 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. The Algorithm 1, therefore, can be efficiently solved. Because the projection (19) preserves the lower-triangular structure, (20) can be efficiently solved using block back substitution, where there is only one factorization cost from the diagonal block, i.e., the reduced block of nominal state matrix

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

Moreover, the reduced state matrices can be repeatedly used when updating the new parameter vector $\mathbf{A}$. In addition, since

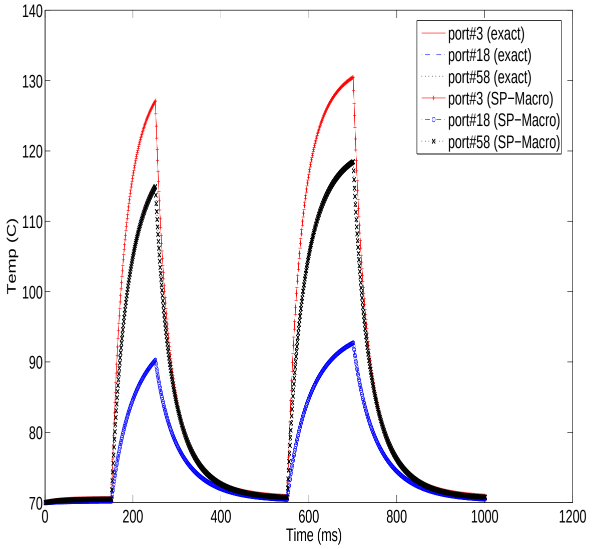

Fig. 6. Transient temperature responses of exact and structured and parameterized macro (SP-Macro) models at port 3, 18, and 58 of layer-1 with step-response input. The macromodels are visually identical to those exact models.

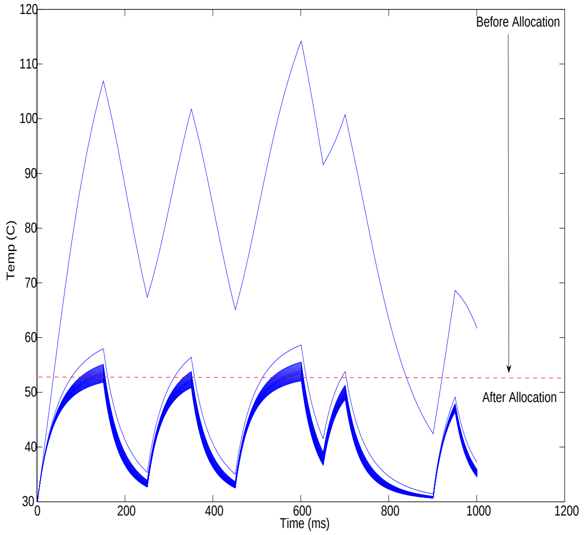

Fig. 7. Iterative optimizations showing the hotspot reduction by thermal-via allocation under the input of transient thermal-power signature at port 32 of layer-1. The ceiling temperature is $52°C$.

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 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 following settings. k_1 (thermal conductive constant) is $100W/m \cdot K$ for silicon and $400W/m \cdot K$ for copper, and k_2 (thermal capacitive constant) is $1.75 \times 10^6 J/m^3 \cdot K$ for silicon and $3.55 \times 10^6 J/m^3 \cdot K$ for copper. The substrate is 500um thick, the device layer is 6um thick and interlayer thickness is 1um thick. 4 silicon layers are used and the thermal-via is assumed to be copper. The unit via area is $2 \times 2um^2$. The overall chip size is $2 \times 2cm^2$, 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 different differently increasing rate.

The power distribution at each title is chosen similarly as [16], where 90% of tiles have power densities from 0 to $2 \times 10^6 W/m^2$, and their clock gating pattern has a period of 500ms, where the power in the standby mode is 5% of the running mode. The other 10% of tiles having power densities from $3 \times 10^6 W/m^2$ to $9 \times 10^6 W/m^2$, and their clock gating pattern has a period of 250ms where the power in the standby mode is 20% of the running mode. In addition, note that a single-input-multi-output (SIMO) [24] is assumed when the port number in B is large.

1) Structured and Parameterized Macromodel: One detailed 3D thermal RC circuit is used to verify the proposed algorithm. It has 4 layers and each layer contains about 1K 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 show 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 during the optimization.

Moreover, Fig. 6 compares the time-domain transient temperature at selected three critical tiles (3, 18, 58) using (22). 16 moments are used for the moment matching. Clearly, the reduced models are visually identical to original ones.

2) Sequential Programming: Furthermore, for the same 3D 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 the ceiling temperature is $52°C$ is meet. In addition, Fig. 8 shows the subgradient optimization procedure after 4 iterations, where the dual problem quickly converges with the primal problem at one normalized value 0.7. Clearly,

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

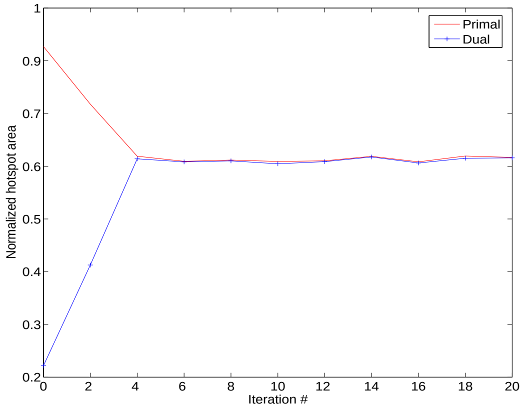

Fig. 8. Convergence of subgradient optimization of primal and dual problems. The hotspot is represented by violation integral normalized to the maximum. α_0 here is set to 0.7.

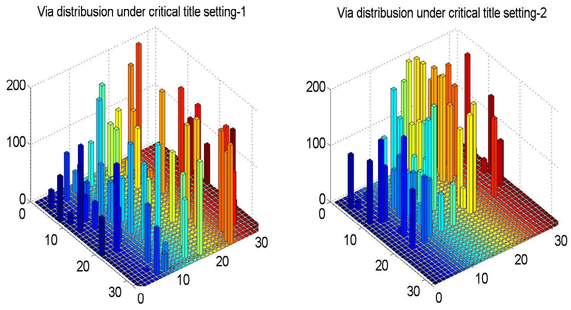

Fig. 9. Thermal-via distribution under two different settings of critical titles.

our sequential programming could effectively minimize the thermal-violation integral efficiently.

3) Thermal-via Allocation: With use of the transient analysis by our macromodel, Fig. 9 further shows the allocated via density distribution for the same 3D thermal circuit above. To study the impact of critical tiles, two power inputs with different clock-gating period are injected. It results in a different set of locations for those critical tiles. Therefore, the allocated via in Fig. 9 (a) shows a little difference with Fig. 9 (b). Therefore, the worst-case power input or workload needs to be assumed that could lead to the worst-case temperature, which accordingly determines the critical tiles.

Moreover, Fig. 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 $150^\circ C$, 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 little 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. The parameterized state equation (17) in the steady-state is used to calculate the transient response and the sensitivity. In addition, as discussed in Section III. A (See Fig. 4), a rectangular area formed by the "steady" maximum temperature and the ceiling temperature is used as the objective instead of the "dynamic" violation integral. Then, the problem is also again solved by a sequential programming.

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 126X (62s versus 7809s) compared to the steady-state analysis. More importantly, due to the use of our accurate figure of merit, the thermal-violation integral, which considers the transient effect, our allocated thermal-via density is much smaller than the one by steady-state analysis under the same targeted ceiling temperature. Because directly solving steady-state equation can not generate the sensitivity for the optimization, the allocated thermal-via of steady-state analysis is based on the reduced macromodel, where the thermal-violation integral is defined by the maximum temperature (See Fig. 4). For a circuit with 32768 tiles, our design reduces 2.04X (11660 versus 23800) thermal vias compared to the steady-state analysis.

VI. CONCLUSIONS AND DISCUSSIONS

The previous thermal-via allocations [16], [17] for 3D IC use 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

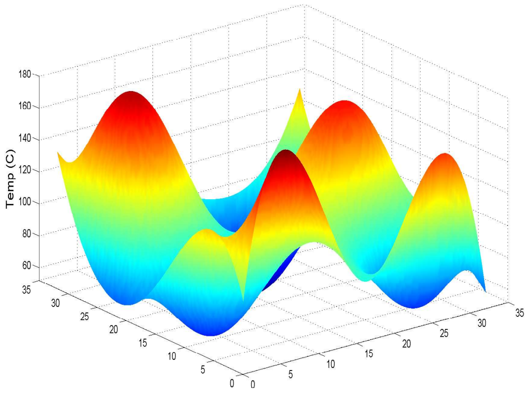

Fig. 10. Steady-state temperature map of top layer (layer-1) before thermal-via allocation.

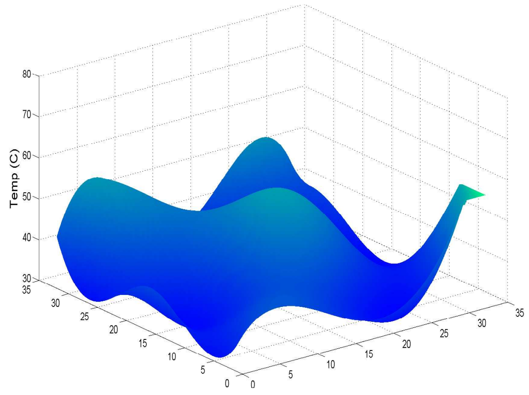

Fig. 11. Steady-state temperature map of top layer (layer-1) after thermal-via allocation using transient temperature profile.

consider the temporally and spatially variant thermal-power input, 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 126X faster to obtain the temperature profile, and reduces the number of thermal vias by 2.04X under the same temperature bound.

Clearly, the proposed structured and parameterized macromodel can be used for a number of integrity-driven physical synthesis. For example, we have recently presented a 3D via planning for simultaneous power and thermal integrity [29], where the 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 used 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 a worst-case temperature profile. As discussed in [19], the worst-case temperature profile and its accordingly related critical titles can be characterized from the thermal-power when the workload is available and guard-land can be used to avoid the under-design. However, it is computationally expensive if not possible to determine the the worst-case temperature profile from all kinds of dynamic workloads. The stochastic characterization approach such as the principal component analysis (PCA) can be applied to find a set of principal temperature and its accordingly related principal 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 non-aligned 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 non-aligned vias by perturbation.

Acknowledgments

This work is funded by NSF CAREER Award CCF-0448534 and UC-MICRO fund from Intel. We deeply appreciate the reviewers for their insightful comments to make the 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*, pp. 602–633, 2001.

[2] S.Das, *Design Automation and Analysis of Three Dimentional Integrated Circuits (Ph. D Thesis)*. Massachusetts Institute of Technology, 2004.

[3] W. Davis and et al., "Demystifying 3D ICs: the pros and cons of going vertical," *IEEE Design and Test of Computers*, pp. 498–510, 2005.

[4] S. Lim, "Physical design for 3D system-on-package: Challenges and opportunities," *IEEE Design and Test of Computers*, pp. 532–539, 2005.

[5] C. C. Teng, Y. K. Cheng, E. Rosenbaum, and S. M. Kang, "iTEM: A temperature-dependent electromigration reliability diagnosis tool," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 882–893, 1997.

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

[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. on Very Large Scale Integration (VLSI) Systems*, pp. 691–700, 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," *Microelectronics Journal*, 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. on Computer Aided Design (ICCAD)*, 2004.

[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. on Computer Aided Design (ICCAD)*, 2005.

[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. on Computer Aided Design (ICCAD)*, 2006.

[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. on Computer Aided Design (ICCAD)*, 2006.

[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. on Very Large Scale Integration (VLSI) Systems*, pp. 501–513, 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. on Computer Aided Design (ICCAD)*, 2003.

[15] J. Cong, J. Wei, and Y. Zhang., “A thermal-driven floorplanning algorithm for 3D ICs,” in *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, 2004.

[16] B. Goplen and S. Sapatnekar, “Thermal via placement in 3D ICs,” in *Proc. Int. Symp. on Physical Design (ISPD)*, 2005.

[17] J. Cong and Y. Zhang., “Thermal via planning for 3D ICs,” in *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, 2005.

[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. on Physical Design (ISPD)*, 2006.

[19] V. Tiwari, D. Singh, S. Rajgopal, G. Mehta, R. Patel, and F. Baez, “Reducing power in high-performance microprocessors,” in *ACM/IEEE Proc. Design Automation Conf. (DAC)*, 1998.

[20] K. Skadron, M. Stan, and et. al., “Temperature-aware microarchitecture,” in *International Symposium on Computer Architecture*, 2006.

[21] W. Liao, L. He, and K. Lepak, “Temperature and supply voltage aware performance and power modeling at microarchitecture level,” *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 1042–1053, 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. on Low Power Electronics and Design (ISLPED)*, 2006.

[23] E.J.Grimme, *Krylov projection methods for model reduction (Ph. D Thesis)*. Univ. of Illinois at Urbana-Champaign, 1997.

[24] A. Odabasioglu, M. Celik, and L. Pileggi, “PRIMA: Passive reduced-order interconnect macro-modeling algorithm,” *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 645–654, 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. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 678–693, 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. on Computer Aided Design (ICCAD)*, 2005.

[27] C. Visweswariah, R. A. Haring, and A. R. Conn, “Noise considerations in circuit optimization,” *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 679–690, 2000.

[28] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*. John Wiley and Sons, 1993.

[29] H. Yu, J. Ho, and L. He, “Simultaneous power and thermal integrity driven via stapling in 3D ICs,” in *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, 2006.