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

Hao Yu, Yiyu Shi, Lei He
Electrical Engineering Dept.
University of California
Los Angeles, CA 90095
{hy255,yshi,lhe}@ee.ucla.edu

Tanay Karnik
Circuit Research
Intel Labs
Hillsboro, OR 97124
tanay.karnik@intel.com

ABSTRACT

All existing methods for thermal-via allocation are based on a steady-state thermal analysis and may lead to excessive number of thermal vias. This paper develops an accurate and efficient thermal-via allocation considering temporally and spatially variant thermal-power. The transient temperature is calculated using macromodel by a structured and parameterized model reduction, *which generates temperature sensitivity with respect to 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. This optimization problem is transformed into a sequence of subproblems by Lagrangian relaxation, and each subproblem is solved by quadratic programming using sensitives from the macromodel. Experiments show that compared to the existing method using 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.*

Categories and Subject Descriptors: B.7.2[Hardware]: Integrated circuits – Design aids

General Terms: Algorithms, Design

Keywords: Thermal Management and Simulation, Model Order Reduction, SQP Optimization

1. INTRODUCTION

3D integration [1, 2] to stack multiple active layered ICs is effective to improve the deep-submicron interconnect performance and increase the transistor packing density. However, due to the increased power density, the heat dissipation is extremely important in 3D-ICs [1]. It is well known that excessively high temperature can significantly degrade interconnect/device reliability and performance [3–5]. One effective heat-removal approach is to use thermal vias to improve the thermal conductivity. Fig. 1 shows the topology of typical 3D-IC designs including the active device layers, thermal-vias, and the substrate.

Because of different workloads and dynamic power management techniques such as clock gating technique extensively used in the modern VLSI design, power has both temporal and spatial variations. A transient thermal-power is the running average of the cycle-accurate power over the scale of the thermal constant [6]. A cycle-accurate micro-architecture level thermal simulation

*This paper is partially supported by NSF CAREER award CCR-0093273/0401682 and Intel. Address comments to lhe@ee.ucla.edu.

ISLPED'06, October 4–6, 2006, Tegernsee, Germany.

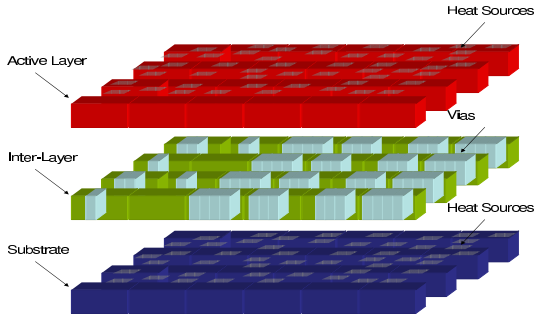

Figure 1: 3D-IC topology including: active device layers, inter-layer dielectrics, vias, and the substrate.

Hotspot [7] has been developed based on a thermal *RC* model to calculate the transient temperature. Assuming steady-state thermal analysis (based on thermal resistance model), thermal-via allocation has been studied during the placement [8] and routing [9]. Because the steady-state analysis ignores the temporal and spatial variations of the transient thermal-power, to obtain a solution without thermal violation, the methods in [8, 9] have to assume a maximum thermal-power *simultaneously* for all regions. Because it is rare for different regions to simultaneously reach their maximum thermal-power, the methods in [8, 9] may lead to excessive number of thermal vias. In addition, [7–9] directly solve the matrix-formed state equation. It can not efficiently calculate the nominal temperature and its sensitivity with respect to the thermal-via density for large sized circuits. The design procedure is either based on iterations [8], or based on an approximated square-root relation [9] between temperature and thermal-vias. It may not converge or may lead to inaccurate results. Therefore, accurate and efficient solutions to calculate temperature and temperature sensitivity should be developed.

In this paper, an accurate yet efficient thermal-via allocation is proposed that considers the temporal and spatial variations of the thermal-power. The transient temperature is calculated using macromodel 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. This optimization problem is transformed into a sequence of subproblems using Lagrangian relaxation, and each subproblem is solved by quadratic programming with the sensitives provided by the macromodel. Experiments show that compared to 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.

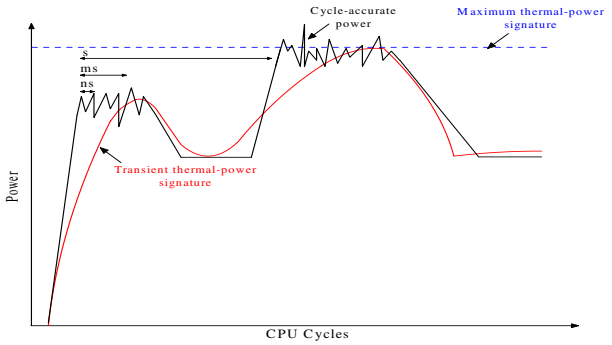

Figure 2: The definitions of cycle-accurate power, transient thermal-power signature, and maximum thermal-power signature at the different scale of time constant.

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

Table 1: Thermal and electrical duality

$N(K)$	number of tiles (critical tiles)
p	number of input/output ports
q	order of reduced models
G_0, C_0	nominal thermal RC state matrices
A_i	via density of ith tile
$x(y)$	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)

Table 2: Notation list

The rest of the paper is organized below. In Section 2, we first present the preliminary for 3D thermal model and analysis. In Section 3, we discuss a structured and parameterized reduction to generate the macromodel. In Section 4, we formulate a nonlinear optimization to accurately allocate the thermal-via driven by the thermal-violation integral. In Section 5, we present experimental results and conclude the paper in Section 6.

2. PRELIMINARY

2.1 Thermal Model

There is a well-known duality between electrical and thermal systems (See Table 1). 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.

The 3D layout can be uniformly discretized into N tiles by the finite difference method. Our design variable 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. In this paper, K critical tiles are assumed to be specified by users. An ith tile has a thermal-via area A_i. Because A_i is related to the thermal-via density ρ_i by $\rho_i = A_i/a$, A_i is used to represent the thermal-via density at ith tile in the sequel. Note that a is the unit area of thermal-via determined by the process.

The equivalent thermal circuit by nodal analysis (NA) in the frequency(s) domain is

$$[G_0 + sC_0 + \sum_{i=1}^{K} A_i(g_i + sc_i)]x(\mathbf{A}, s) = Bu(s)$$
$$y(\mathbf{A}, s) = L^T x(\mathbf{A}, s), \quad (1)$$

where $\mathbf{A} = [A_1, ..., A_K]$ is a parameter-vector of thermal-via density. Note that G_0 and C_0 ($\in R^{N\times N}$) are conductive and capacitive matrices of discretized thermal networks, and $\sum_{i=1}^{K} A_i g_i$ and $\sum_{i=1}^{K} A_i c_i$ are conductive and capacitive matrices of thermal vias, respectively. In addition, $x(\mathbf{A}, s)$ ($\in R^N$) is the state variable of node temperatures, B ($\in R^{N\times p}$) is the adjacent matrix to select input u, and L ($\in R^{N\times p}$) is the adjacent matrix to select output y. The notations are summarized in Table 2.

The thermal-via is 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 at two between two vertical-adjacent layers, its insertion matrix is

$$X(k,l) = X(l,k) = \begin{cases} -1 & \text{if } k = m,\ l = n \\ \sum_l |X(k,l)| & \text{if } k = l \\ 0 & \text{else} \end{cases}. \quad (2)$$

Accordingly, we have $g_i = (k_1/t)X_i$ and $c_i = (k_2 t)X_i$, where k_1 and k_2 are thermal conductive/capacitive constants of the thermal-via, w and t are the width and thickness of the thermal-via.

Moreover, note that u ($\in R^{p\times 1}$) is the current source to model the thermal-power input. There are several types of thermal-power as defined in [6]. A *thermal power* is defined by 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). When the set of architectural model/constraints and the particular instruction sets and working loads driving the chip are available, a *transient thermal-power signature* can be further defined as the thermal power with a worst-case trace input [6]. In addition, a constant *maximum thermal-power signature* is defined as the maximum of the transient thermal-power signature. Fig. 2 illustrates differences of these power definitions.

2.2 Thermal Analysis

The direct solution in [7–9] is not efficient to solve (1) for large sized circuits. Similar to the macromodeling for the electrical RC network, moment matching based model order reduction can be used to obtain a compact thermal RC model, which not only has a smaller matrix size but also preserves the dominant system response. The existing macromodeling approach from electrical analysis is mainly based on the subspace projection [10] by expanding the system equation (1) at some frequency points. After projection, an order reduced state equation can be obtained with preserved low-order moments to represent the dominant response of the original system.

To further obtain the sensitivity information, the parametrized moments [11] can be obtained by expanding (1) 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 [12] 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 [10–12] apply a flat projection during the reduction. The reduced state matrices and 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.

3. 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, and 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.

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:

$$x(\mathbf{A}, s) = \sum_{i_1}^{\infty} \cdots \sum_{i_K}^{\infty} x_{1,...,K}^{(i_1+...+i_K)}(s)(\delta A_1)^{i_1} \cdots (\delta A_K)^{i_K}. \quad (3)$$

This is similar to the method in [12] modeling variations for the electrical system. Substituting (3) in (1), explicitly matching the moment for each A_i up to the second-order, we can reformulate (1) into an augmented parameterized state equation:

$$(G_{ap} + sC_{ap})x_{ap} = B_{ap}u(t), \quad y_{ap} = L_{ap}^T x_{ap}, \quad (4)$$

with

$$G_{ap} = \begin{bmatrix} G_0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ A_1 g_1 & G_0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ A_K g_K & 0 & \dots & G_0 & 0 & 0 & \dots & 0 \\ 0 & A_1 g_1 & 0 & \dots & G_0 & 0 & \dots & 0 \\ 0 & A_2 g_2 & A_1 g_1 & 0 & \dots & G_0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_k g_K & \dots & 0 & \dots & G_0 \end{bmatrix} \quad (5)$$

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 (4) 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 (4) is large, its order needs to be reduced using projection with preserved moments (of s) up to q-th order. A flat projection matrix V can be constructed recursively using Arnoldi method [12]. However, directly projecting (4) by V leads to a reduced macromodel losing the lower-triangular block structure of G_{ap} and C_{ap}. As a result, $y^{(0)}$, $y^{(1)}$ and $y^{(2)}$ are coupled with each other.

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

$$\mathcal{V} = diag[V_0, \underbrace{V_1, ..., V_K}_{K}, \underbrace{V_{K+1}, ..., V_{K^2}}_{K^2}], \quad (6)$$

by partitioning V according to the dimension of $x^{(0)}$, $x^{(1)}$ and $x^{(2)}$. As a result, the order-reduced state matrices

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

Because $V \subseteq \mathcal{V}$, a q-th ordered projection by $\mathcal{V}$ still preserves at least q moments according to [13].

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

$$\begin{aligned} (\widetilde{G}_{ap} + \frac{1}{h}\widetilde{C}_{ap})\widetilde{x}_{ap}(t) &= \frac{1}{h}\widetilde{C}_{ap}\widetilde{x}_{ap}(t-h) + \widetilde{B}_{ap}u(t) \\ \widetilde{y}_{ap}(t) &= \widetilde{L}_{ap}^T \widetilde{x}_{ap}(t). \end{aligned} \quad (7)$$

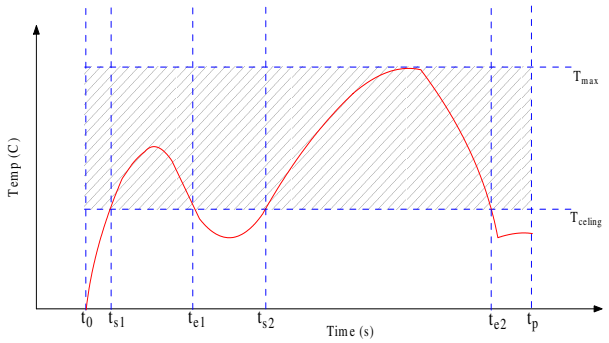

Figure 3: Figure of merit using thermal-violation integral with defined ceiling temperature under an input of transient thermal-power signature.

where

$$\widetilde{G}_{ap} = \begin{bmatrix} \widetilde{G}_0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ A_1 \widetilde{g}_1 & \widetilde{G}_0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ A_K g_K & 0 & \dots & \widetilde{G}_0 & 0 & 0 & \dots & 0 \\ 0 & A_1 \widetilde{g}_1 & 0 & \dots & \widetilde{G}_0 & 0 & \dots & 0 \\ 0 & A_2 g_2 & A_1 \widetilde{g}_1 & 0 & \dots & \widetilde{G}_0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_k \widetilde{g}_K & \dots & 0 & \dots & \widetilde{G}_0 \end{bmatrix} \quad (8)$$

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

$$\widetilde{y}(\mathbf{A}, t) = \widetilde{y}^{(0)}(\mathbf{A}, t) + \widetilde{y}^{(1)}(\mathbf{A}, t) + \widetilde{y}^{(2)}(\mathbf{A}, t), \quad (9)$$

A thermal-via planning based on the accurate yet efficient transient simulation with $\widetilde{y}(\mathbf{A}, t)$ can be consequently design. Note that as the reduced system still has the lower-triangular structure, (7) 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.

4. THERMAL-VIA ALLOCATION

In this Section, an accurate figure of merit, thermal-violation integral is first defined to consider the transient temperature profile. A thermal-via allocation can consequently be formulated as a nonlinear optimization problem, which is relaxed and solved by a sequence of quadratic programmings with use of sensitivities provided from the structured and parameterized macromodel.

4.1 Thermal-Violation Integral

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

$$\begin{aligned} f_i(\mathbf{A}) &= \int_{t_0}^{t_p} max[\widetilde{y}(\mathbf{A}, t), T_{ceiling}]dt \\ &= \int_{t_s}^{t_e} [\widetilde{y}(\mathbf{A}, t) - T_{ceiling}]dt, \end{aligned} \quad (10)$$

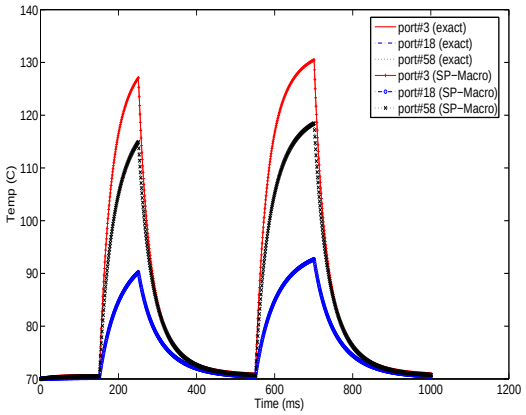

Figure 4: Transient temperature responses of exact and structured and parameterized macro (SP-Macro) models at port 3, 18, and 58 of layer-1 with step-response input. The macromodels are visually identical to those exact models.

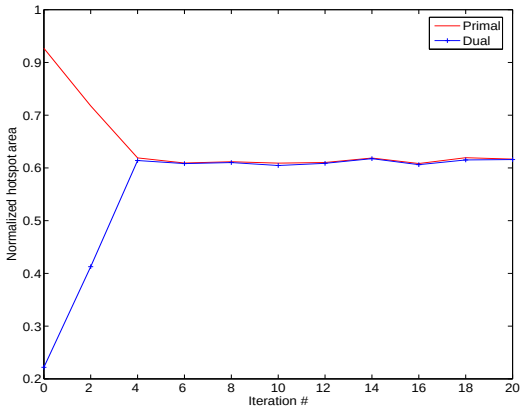

Figure 5: Convergence of subgradient optimization of primal and dual problems. The hotspot is represented by violation integral normalized to the maximum. α_0 here is set to 0.7.

where $\mathbf{A} = [A_1, ..., A_K]$ is parameter vector of thermal-via density, t_0 and t_p define time-period, and the interval $[t_s, t_e]$ is determined by comparing

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

which can contain multiple intervals. As shown in Fig. 3, 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 [14].

Moreover, the figure of merit for a group of P critical tiles in the entire circuit is

$$f(\mathbf{A}) = \sum_{i=1}^{P} f_i(\mathbf{A}). \tag{11}$$

It is called *total thermal-violation integral*. The total thermal-violation integral is used as an accurate objective function in the sequel to be minimized by allocating thermal vias.

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. 3). 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.

4.2 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 optimization problem is formulated as

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

$$s.t. \quad \sum_{i=1}^{K} A_i \le A_{max}, \tag{12}$$

$$0 \le A_i \le (A_i)_{max}, (i = 1, ..., K). \tag{13}$$

where the constraint (12) is a *global constraint* implying that the total thermal-via density is limited by the A_{max}, and the constraint (13) is a *local constraint* implying that the local thermal-via density at ith tile is limited by $(A_i)_{max}$. Moreover, to compute $f(\mathbf{A})$, t is discretized into finite intervals and Problem 1 becomes semi-definite [14], which can be further solved using Lagrangian relaxation.

Using matrix $\mathbf{U}$ $(\in R^{(K+1)\times(K)})$

$$\mathbf{U} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 1 \\ 1 & 1 & \dots & 1 \end{bmatrix}, \tag{14}$$

the constraints (12) and (13) become

$$\mathbf{UA} \le \mathbf{A}_{max}, \tag{15}$$

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 is used to transform the original problem into a sequence of subproblems.

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:

$$L(\mathbf{A}, \lambda) = f(\mathbf{A}) + \lambda \cdot \mathbf{h}(\mathbf{A}) \tag{16}$$

where

$$\mathbf{h}(\mathbf{A}) = \mathbf{UA} - \mathbf{A}_{max}. \tag{17}$$

This relaxed problem can be transformed into a sequential subproblems by subgradient optimization [15]. At each iteration, each subproblem is constructed from a quadratic approximation of the nonlinear objective function, and a linearization of the constraints about the solutions from previous iteration. The optimization terminates when the convergence criterion is achieved. This called as *sequential quadratic programming* (SQP) [15].

Expanding $f(\mathbf{A})$ and $h(\mathbf{A})$ with respect to $\mathbf{A}$ up to the second-order, an approximated equivalent subproblem is

$$min \quad \nabla f(\mathbf{A})^T \delta\mathbf{A} + \frac{1}{2}\delta\mathbf{A}^T H \delta\mathbf{A} \tag{18}$$

$$s.t. \quad \nabla\mathbf{h}(\mathbf{A}) \cdot \delta\mathbf{A} \le \mathbf{h}(\mathbf{A}). \tag{19}$$

(19) can be solved by the standard quadratic programming, where

$$\nabla f = \int_0^{t_p} \widetilde{y}^{(1)} dt, \quad \nabla h = const.$$

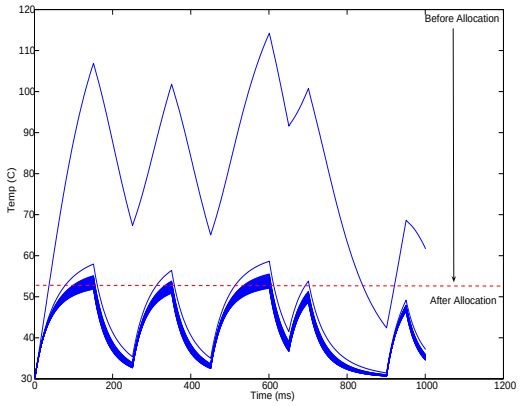

Figure 6: Iterative optimizations showing the hotspot reduction by thermal-via allocation under the input of transient thermal-power signature at port 32 of layer-1. The ceiling temperature is $52^{\circ}C$.

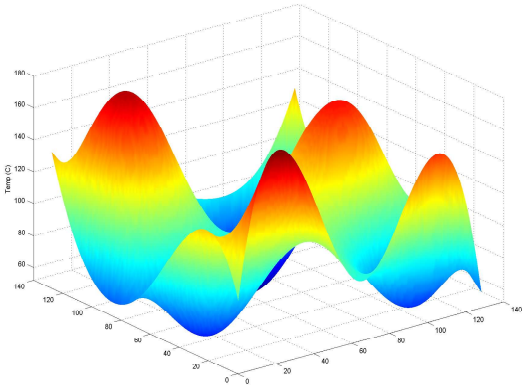

Figure 7: Steady-state temperature map of top layer (layer-1) before thermal-via allocation.

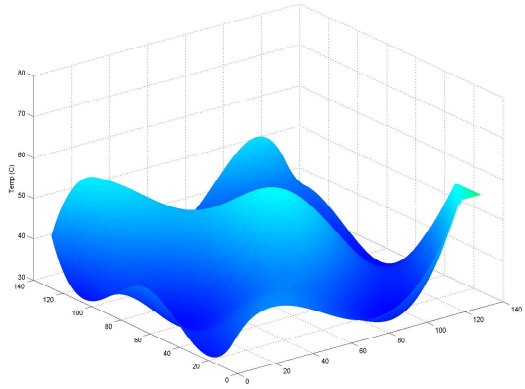

Figure 8: Steady-state temperature map of top layer (layer-1) after thermal-via allocation using transient temperature profile.

are first-order sensitivities, and

$$H = \begin{bmatrix} \int_0^{t_p} \widetilde{y}_{1,1}^{(2)} dt & \cdots & \int_0^{t_p} \widetilde{y}_{1,K}^{(2)} dt \\ \vdots & \ddots & \vdots \\ \int_0^{t_p} \widetilde{y}_{K,1}^{(2)} dt & \cdots & \int_0^{t_p} \widetilde{y}_{K,K}^{(2)} dt \end{bmatrix}$$

is the Hessian matrix composed by the second-order sensitivities. Both the first and second order sensitivities can be efficiently solved by (7) independently.

The sequential subgradient optimization procedure is outlined in Algorithm 1, where α_k is the step size usually determined through a geometric regression [15]. Note that because the projection (6) preserves the block structure, the reduced state matrices can be repeatedly used when updating the new parameter vector $\mathbf{A}$. Therefore, there is only one reduction needed. In addition, since the reduced model is much smaller than the original one, and the factorization cost only comes from the nominal blocks in diagonal, its nominal value and sensitivities can be efficiently solved by the back-substitution of (7). Therefore, the optimization procedure in Algorithm 1 is computationally efficient.

Algorithm 1 Subgradient Optimization using Structured Parameterized Macromodel

Initialize: $(\mathbf{A}_0, \alpha_0, \lambda_0, H_0, k)$;
Solve: $\widetilde{y}_0$ using (7);
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 (7) with updated macromodel;
  $k = k + 1$;
end while

5. EXPERIMENTS

Our structured and parametrized macromodeling (SP-Macro) and thermal-via allocation are both implemented in MATLAB, 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. A random power distribution at each node is used. 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.

A detailed 3D thermal RC circuit is used to verify the proposed algorithm. It has 4 layers and each layer contains about 10K 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 10 to 100. 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 SP-macro and original models, Fig. 4 compares the time-domain transient temperature at selected three critical tiles (3,

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

Table 3: 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.

18, 58) using (9). 16 moments are used for the moment matching. The reduced models are visually identical to original ones. Fig. 5 shows the subgradient optimization procedure after few iterations, where the dual problem converges with the primal problem. The ceiling temperature is $52^\circ C$ and, the transient temperature at one port is cooled down to the ceiling point as shown in Fig. 6. Clearly, the gradient approach greedily minimizes the thermal-violation integral. Fig. 7 and 8 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. 7. In contrast, the allocation results in a cooled temperature profile that closely approaches the ceiling temperature as shown in Fig. 8.

Table 3 further analyzes the runtime scalability and allocated thermal-via density by the proposed method and the direct steady-state analysis. Because directly solving steady-state equation needs to handle large sized matrix, it has a long runtime and uses a lot of memory. 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. 3). For a circuit with 32768 tiles, our design reduces 2.04X (11660 versus 23800) thermal vias compared to the steady-state analysis.

6. CONCLUSIONS

An accurate yet efficient thermal-via allocation is proposed for the thermal-aware design of 3D ICs. The previous thermal-via allocations [8,9] use the direct steady-state analysis and ignore the temporal and spatial variations of the thermal-power. They are inefficient to generate the nominal temperature and its sensitives for large sized circuits. More importantly, they result in a design with excessive number of thermal vias.

In this paper, to consider the temporally and spatially variant thermal-power input, a structured and parameterized model order reduction is used to obtain a macromodel, which can efficiently provide the transient nominal temperature and its sensitives to thermal-via densities. A thermal-violation integral of the transient temperature is then defined to accurately capture the thermal violation, and a nonlinear optimization is formulated to minimize the thermal-violation integral. In addition, using parameterized sensitivities provided from the macromodel, the relaxed subproblems of the formulated problem are efficiently solved by a sequence of quadratic programming, where the reduced macromodel can be repeatedly used during the gradient search. Clearly, the proposed structured and parameterized macromodel can be used for a number of integrity-driven physical synthesis.

7. 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] 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.

[3] C. C. Teng, Y. K. Cheng, E. Rosenbaum, and S. M. Kang, "iTEM: A temperature-dependent electromigration reliability diagnosis tool," *IEEE Trans. on CAD*, pp. 882–893, 1997.

[4] K.Banerjee, A.Mehrotra, A.Sangiovanni-Vincentelli, and C. Hu, "On thermal effects in deep sub-micron VLSI interconnects," in *ACM/IEEE DAC*, 1999.

[5] W. Huang, E. Humenay, K. Skadron, and M. R. Stan, "The need for a full chip and package thermal model for thermally optimized IC designs," in *ACM/IEEE ISLPED*, 2005.

[6] V. Tiwari, D. Singh, S. Rajgopal, G. Mehta, R. Patel, and F. Baez, "Reducing power in high-performance microprocessors," in *ACM/IEEE DAC*, 1998.

[7] 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.

[8] B. Goplen and S. Sapatnekar, "Thermal via placement in 3D ICs," in *ACM ISPD*, 2005.

[9] J. Cong and Y. Zhang., "Thermal via planning for 3D ICs," in *IEEE/ACM ICCAD*, 2005.

[10] A. Odabasioglu, M. Celik, and L. Pileggi, "PRIMA: Passive reduced-order interconnect macro-modeling algorithm," *IEEE Trans. on CAD*, pp. 645–654, 1998.

[11] 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 CAD*, pp. 678–693, 2004.

[12] 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 *IEEE/ACM ICCAD*, 2005.

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

[14] C. Visweswariah, R. A. Haring, and A. R. Conn, "Noise considerations in circuit optimization," *IEEE Trans. on CAD*, pp. 679–690, 2000.

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