Simultaneous Power and Thermal Integrity Driven Via Stapling in 3D ICs *

Hao Yu, Joanna Ho, Lei He
EE Dept., University of California
Los Angeles, CA 90095
{hy255,jho,lhe}@ee.ucla.edu

ABSTRACT

The existing work on via-stapling in 3D integrated circuits optimizes power and thermal integrity separately and uses steady-state thermal analysis. This paper presents the first in-depth study on simultaneous power and thermal integrity driven via-stapling in 3D design. The transient temperature and supply voltage violations are calculated by a structured and parameterized model reduction, which also generates parameterized temperature and voltage violation sensitivities with respect to the via pattern and density. Using parameterized sensitivities, an efficient yet effective greedy optimization is presented to optimize power and thermal integrity simultaneously. Experiments with two active device layers show that compared to sequential power and thermal optimization using steady-state thermal analysis, sequential optimization using transient thermal analysis reduces non-signal vias by on average 11.5%, and simultaneous optimization using transient thermal analysis reduces non-signal vias by on average 34%. The via reduction would be higher for the 3D design with more device layers.

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

General Terms: Algorithms, Design

Keywords: Thermal Modeling and Management, Structured and Parameterized Model Order Reduction

1. INTRODUCTION

Compared to the conventional two dimensional (2D) integration with one active device layer, the three dimensional (3D) integration with multiple active layers is effective to increase integration level and improve performance [1–3]. However, 3D integration also creates such challenges as power and thermal integrity, more difficult to deal with than those in 2D integration. Fig. 1 illustrates a typical 3D stacking of multiple active layers inside a single package. Power is supplied from the bottom, the power and ground planes in the package. C4 bumps connect the power and ground planes to the active device layers, and through-vias that could be called as "power/ground vias" are used to carry power supply between active layers. Due to the strong electromagnetic coupling between the package and the power delivery system (PDS) [4], it is a critical to optimize power integrity in 3D ICs [5]. As shown by the PDS designs in 2D ICs [6–8], stapling power/ground vias reduces the loop inductance of power/ground planes, and hence reduces the SSN (simultaneously switching noise) in package. However, there is no in-depth study on design automation of PDS in 3D ICs.

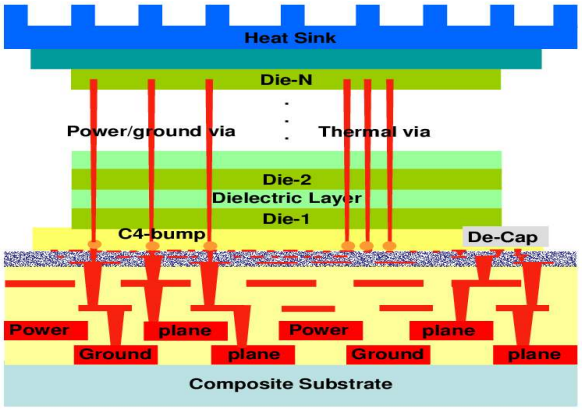

Figure 1: A typical 3D stacking with non-signal through-vias.

Moreover, because of 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 [9–11]. Since vias are good thermal conductors as well, adding through-vias as "thermal vias" between device layers is effective to remove heat [12, 13]. A heat-sink is needed when the chip power is beyond 25Watt. It is usually placed on the top of device layers and serves as the primary heat-removal path. As shown in Fig. 1, thermal vias can be inserted to more effectively remove heat from bottom device layers to the heat sink.

We call power/ground and thermal vias as non-signal vias. The number of these vias could be very large. For example, more than 10^4 thermal vias were used in [12, 13]. Because large numbers of non-signal vias introduce congestion for signal vias, planning non-signal vias in 3D ICs becomes a need. Existing via-stapling [6–8, 12, 13] implicitly optimizes power and thermal integrity *separately*, where power/ground vias are inserted to satisfy power integrity constraints, and thermal vias are inserted to satisfy thermal integrity constraints. These vias are stapled according to the distribution of maximum temperature and voltage violations. Because maximum voltage violations are often located differently from maximum temperature violations, the resulting vias can have quite different patterns. Fig. 2 shows the spatial distribution of the normalized temperature and voltage violations for a typical 3D design before stapling vias. Since the combined

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

ICCAD'06, November 5–9, 2006, San Jose, CA.

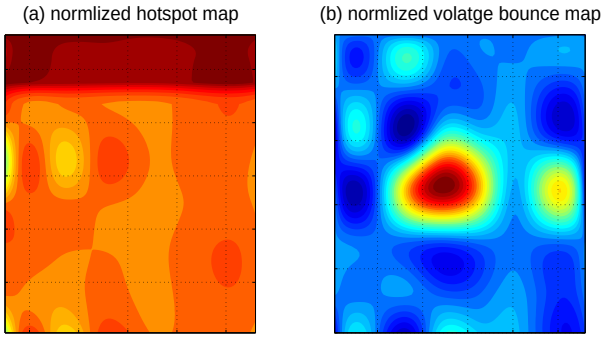

Figure 2: (a) is the temperature violation distribution map and (b) is the supply voltage violation distribution map. Both are before the via stapling. The region with the dark(red)-color has the maximum temperature or voltage value. The two distributions are different.

power/ground planes work as a cavity resonator, large voltage violation can be found often in the center of the planes [4,6]. On the other hand, the thermal hotspots are those regions close to the heating sources on the device layer and they may spread more uniformly for a thermal aware 3D placement and routing [12,13]. As a result, it leads to two different via-stapling patterns: the vias tend to be stapled in the center for the power integrity but stapled uniformly across the plane for the thermal integrity. Because stapling vias in such a *sequential* fashion ignores that non-signal vias could be used to minimize both the temperature and voltage violation, it may result in an over-design. Furthermore, to obtain a valid solution, the existing thermal-via planning [12,13] assumes a steady-state thermal analysis with the maximum thermal-power as inputs. Since it is rare if not impossible for different regions to simultaneously reach their maximum thermal-power, the assumption of steady-state analysis may also lead to excessive numbers of vias. Therefore, it urges us to provide a stapling method using transient analysis to find the minimum number of non-signal vias, such that both power integrity in P/G planes and thermal integrity at device layers can be satisfied.

In this paper, we formulate and solve the 3D via-stapling problem to minimize non-signal vias subject to power and thermal integrity simultaneously. We apply transient models for temperature and supply voltage noise. As shown by experiments, for a sequential power and thermal optimization, using transient analysis reduces stapled non-signal vias by an average of 11.5% compared to using steady-state analysis. Moreover, our simultaneous optimization of power and thermal integrity by transient analysis reduces on average 34% non-signal vias compared to the sequential optimization by steady-state analysis.

The rest of the paper is organized as follows. We discuss the model and formulate the level based via-stapling problem in Section 2. We introduce a fast integrity analysis in Section 3, and develop an efficient algorithm in Section 4. We present experiments in Section 5 and conclude the paper in Section 6.

2. MODELING AND PROBLEM FORMULATION

2.1 Distributed Circuit Model

We model the 3D design by two parts: distributed thermal-RC circuit for thermal integrity and electrical-RLC circuit for power integrity. 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 through a pair of thermal resistance and capacitance driven by the current source, which in turn models the power dissipation.

The 3D layout of each silicon device layer and power/ground plane can be uniformly discretized by the finite difference method.

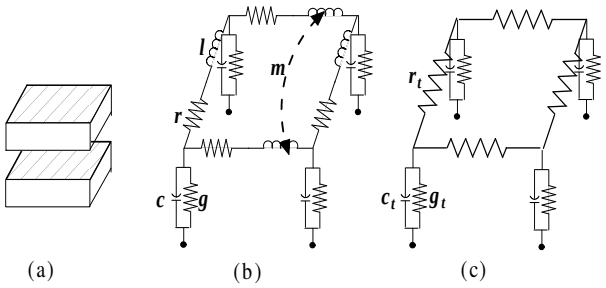

Figure 3: (a) is a unit tile, (b) is its equivalent RLC-cell, and (c) is its equivalent RC-cell.

Temperature	Voltage state variables ($x(t)$)
Input Thermal-Power	Input Current sources ($\mathbf{I}(t)$)
Thermal conductance	Electrical conductance (G)
Thermal capacitance	Electrical capacitance (C)

Table 1: Thermal and electrical duality

As shown in Fig. 3, each discretized tile can be represented by an RC-cell and an RLC-cell to construct distributed thermal-RC and electrical-RLC circuits, respectively. The via or C4 bump is modeled by a lumped RC pair for both thermal and electrical circuits. Moreover, we use partial inductance [14] to consider the magnetic coupling in the electrical-RLC circuit.

2.2 Thermal Analysis

According to [15], a *transient thermal-power* is the running average of the cycle-accurate (often in the range of ns) power over the thermal time constant (often in the range of ms). When working loads are known, a constant *maximum thermal-power* can be calculated as the maximum of the transient thermal-power, and should be calculated for each region of the chip. Fig. 4 illustrates differences of these power definitions.

Assuming steady-state thermal analysis (based on thermal resistance model), thermal-via allocation has been studied for the placement [12] and routing [13]. Because the steady-state analysis ignores the temporal and spatial variations of the transient thermal-power, the methods in [12, 13] have to assume the maximum thermal-power *simultaneously* for all tiles in the integrated circuit to obtain a solution without thermal violation. Because it is rare if not impossible for different tiles to simultaneously reach their maximum thermal-power, the methods in [12, 13] may lead to excessive number of thermal vias. In addition, they directly solve the matrix-formed state equation. It can not efficiently calculate the nominal temperature and its sensitivity with respect to the via density for large scale designs. The design procedure is either based on iterations [12], or based on an approximated square-root relation [13] between temperature and thermal-vias. It may not converge or may lead to inaccurate results. Therefore, accurate and efficient solutions to transient temperature and temperature sensitivity should be developed and are discussed in Section 3.

2.3 Level based Via-Stapling Problem

During the early planning stage, vias are vertically stapled between each pair of aligned tiles in adjacent active device layers. The stapling may have different patterns, which can be described by *levels*:

DEFINITION 1. *Assuming each layer is discretized into N tiles, a level-i ($i = 1, ..., K$) stapling pattern is to hierarchically and symmetrically select 4^i tiles from each layer by subdivision, and staple vias for each pair of aligned tiles in adjacent layers.*

The levelized via-stapling patterns are shown in Fig. 5. A level-0 pattern means stapling vias in the center tile, and a level-1 pattern means stapling vias in each of the 4 partitions. As we can

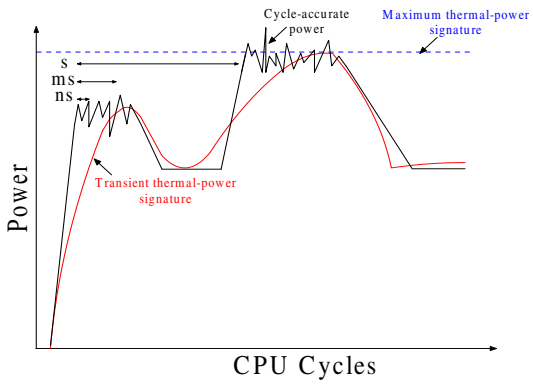

Figure 4: The definitions of cycle-accurate power, transient thermal-power signature, and maximum thermal-power signature at the different scale of time constant.

see, the via-pattern becomes more uniform as the level increases. Multiple levels can be used simultaneously for a design, and we assume that the maximum level number K is provided by users.

Considering total K levels of via-stapling patterns, our design freedom is the via-density (number of vias in one tile) for each level. If we define a via-density vector

$$\mathbf{D} = [D_1, D_2, \cdots, D_K], \tag{1}$$

each level is then associated with a via density D_i to be decided during the via-stapling.

Accordingly we have the following problem formulation:

FORMULATION 1. *Given the allowed maximum voltage violation V_{max} in P/G planes and the allowed maximum temperature T_{max} in device layers, the via-stapling problem is to minimize the total via number, such that the temperature is smaller than T_{max} and the voltage violation is smaller than V_{max}.*

This simultaneous power and thermal integrity driven problem can be represented by

$$\begin{aligned} &min \sum_{i=1}^{K} n_i D_i \\ &s.t.\ \max(V) \le V_{max}, \quad \max(T) \le T_{max} \\ &and\ D_{min} \le D_i \le D_{max} \end{aligned} \tag{2}$$

where n_i is the number of tiles to be stapled in level i. D_{max} is decided by the signal-via routing congestion and D_{min} is decided by the current density (reliability). The key to solving (2) is an efficient yet accurate transient integrity analysis. Such an analysis is presented in Section 3.

3. INTEGRITY ANALYSIS

3.1 Parameterized Description

Because the distributed thermal-RC circuit is a simpler case for the electrical-RLC circuit, we only present the RLC circuit for modeling and algorithm description unless stated otherwise.

Note that the via density D_i at one tile is related to the via area A_i in the tile by $D_i = A_i/(n_i \cdot a)$. a is the unit area of via determined by the processing technology. Because conductance and capacitance values are all proportional to the area A, they are implicitly proportional to the via density D as well. Therefore, the electrical-RLC circuit with the parameterized number of vias can be formulated in MNA (modified nodal analysis) in frequency (s) domain:

$$\begin{aligned} [\mathcal{G}_0 + s\mathcal{C}_0 + \sum_{i=1}^{K} D_i(g_i + sc_i)]x(\mathbf{D}, s) &= \mathcal{B}\mathbf{I}(s) \\ y(\mathbf{D}, s) &= \mathcal{B}^T x(\mathbf{D}, s) \end{aligned} \tag{3}$$

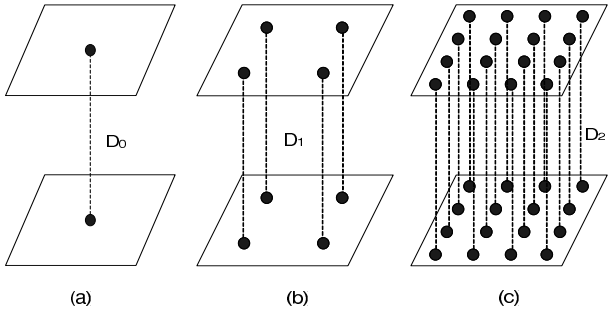

Figure 5: The levelized via-stapling (a) level = 0, (b) level = 1 and (c) level = 2 to describe via patterns. Vias are stapled between each pair of adjacent layers.

$x(y)\ (\in R^{M\times 1})$	State variable (at output)
$\mathbf{v}_n\ (\in R^{N_v\times 1})$	Nodal voltage variables
$\mathbf{i}_l\ (\in R^{N_l\times 1})$	Inductive-branch current variables
$G\ (\in R^{N_v\times N_v})$	Nominal conductance matrix
$C\ (\in R^{N_v\times N_v})$	Nominal capacitance matrix
$L\ (\in R^{N_l\times N_l})$	Nominal inductance matrix
$\mathbf{E}_i\ (\in R^{N_v\times p})$	Input/output incident matrix
$\mathbf{E}_l\ (\in R^{N_v\times N_l})$	Inductive incident matrices
$D_i\ (\in R^{M\times M})$	Parameterized via-density

Table 2: Notation list for system equation (3). Note that $M = N_v + N_l$.

with

$$x(s) = \begin{bmatrix} \mathbf{v}_n \\ \mathbf{i}_l \end{bmatrix}, \mathcal{B} = \begin{bmatrix} \mathbf{E}_i \\ 0 \end{bmatrix}.$$

and

$$\mathcal{G}_0 = \begin{bmatrix} G & \mathbf{E}_l \\ -\mathbf{E}_l^T & 0 \end{bmatrix}, \mathcal{C}_0 = \begin{bmatrix} C & 0 \\ 0 & L \end{bmatrix} \tag{4}$$

All notations in (4) are summarized in Table 2. Note that $\mathcal{B}$ is the adjacent matrix composed by $\mathbf{E}_i$. It describes p_i inputs and p_o critical nodes to be probed, both provided by users.

To mathematically describe adding vias in a circuit equation, an *insertion (adjacent) matrix* X is introduced. For a level-i stapling, adding vias between tiles m and n results in:

$$\begin{aligned} &X_i(k,l) = X_i(l,k) \\ &= \begin{cases} -1 & \text{if } k = m,\ l = n \text{ and } k \ne l \\ \sum_l |X_i(k,l)| & \text{if } k = l \\ 0 & \text{else} \end{cases} \end{aligned} \tag{5}$$

where

$$k, l \in 1, 2, \cdots, 4^i \quad i \in 1, 2, \cdots, K.$$

Then, the unit conductance and capacitance matrices for the level-i stapling are:

$$g_i = gX_i, \quad c_i = cX_i$$

where g and c are conductance and capacitance for the via with unit area a.

The vias are then added to the nominal $\mathcal{G}_0$ and $\mathcal{C}_0$ as a perturbed adjustment:

$$\mathcal{G} = \mathcal{G}_0 + \sum_{i=1}^{K} D_i g_i, \quad \mathcal{C} = \mathcal{C}_0 + \sum_{i=1}^{K} D_i c_i.$$

where $\mathcal{G}$ and $\mathcal{C}$ are the adjusted state matrices including the perturbation from vias.

Note that there are the following differences between thermal-RC and electrical-RLC circuits in MNA. For a thermal-RC circuit, (4) becomes

$$\mathcal{G}_0 = G, \quad \mathcal{C}_0 = C \tag{6}$$

where G and C has larger RC values and results in a larger time-constant than an electrical-RLC circuit does. In addition, the input $\mathbf{I}(s)$ for the thermal-RC circuit stands for thermal-power. In contrast, $\mathbf{I}(s)$ stands for switching-current for electrical-RLC circuits. Moreover, output y at the selected critical nodes is temperature T and voltage V for the thermal-RC and electrical-RLC circuit, respectively.

3.2 Macromodel by Moment Matching

It is inefficient to directly solve (3) for large scale designs. Macromodeling technique based on moment matching can obtain compact models for large distributed thermal-RC and electrical-RLC circuits. We first review the existing flat reduction in this subsection, and then introduce a structured and parameterized reduction in Section 3.3.

To build macromodel by moment matching, we first define moment generation matrices (expanded at s_0) as

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

We then obtain the following projection matrix Q that spans the qth block Krylov subspace by

$$\mathcal{K}(\mathcal{A}, \mathcal{R}, q) = \{\mathcal{A}, \mathcal{A}\mathcal{R}, \cdots \mathcal{A}^{q-1}\mathcal{R}\} \subseteq Q$$

which can be constructed by a block Arnoldi method [16]. As a result, a reduced system can be obtained by projection:

$$\hat{\mathcal{G}} = Q^T\mathcal{G}Q, \quad \hat{\mathcal{C}} = Q^T\mathcal{C}Q, \quad \hat{\mathcal{B}} = V^T\hat{\mathcal{B}},$$

with

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

Note that $\hat{H}$ accurately approximate the original H by matching the first q block moments expanded at s_0 [17]. This procedure can be applied for both thermal-RC and electrical-RLC circuits.

The above macromodeling, however, can only generate nominal values. By expanding (3) at design parameter points such as the via density at different levels, parametrized moments, i.e., the sensitivities of temperature and voltage with respect to the via density can be obtained. Because the parameterized moments have coupled frequency and parameter variables [18], their dimensions grow exponentially and they may not be used in practice. This is improved in [19] by separately expanding moments of design parameters from those of frequency. However, [18, 19] apply a flat projection during the reduction and destroy the matrix structure. As a result, the nominal values and sensitivities after reduction can not be separated easily.

3.3 Structured and Parameterized Reduction

In this paper, we show that nominal values and their sensitivities can be obtained separately by a structured and parameterized reduction. Because the sensitivity is large with respect to the frequency change but small with respect to the design parameter perturbation, the state variable $x(\mathbf{D}, s)$ can be approximated by Taylor expansion:

$$x(\mathbf{D}, s) = \sum_{i_1}^{\infty} \cdots \sum_{i_K}^{\infty} x_{1,\ldots,K}^{(i_1+\ldots+i_K)}(s)(\delta D_1)^{i_1} \cdots (\delta D_K)^{i_K}. \tag{7}$$

This is similar to the method in [19] to deal with process variations. Substitute (7) in (3), and explicitly match the moment for each D_i to the first-order, (3) can be reformulated into an augmented parameterized state equation:

$$(\mathcal{G}_{ap} + s\mathcal{C}_{ap})x_{ap} = \mathcal{B}_{ap}\mathbf{I}(s), \quad y_{ap} = \mathcal{L}_{ap}x_{ap}, \tag{8}$$

with

$$\mathcal{G}_{ap} = \begin{bmatrix} \mathcal{G}_0 & 0 & \ldots & 0 \\ D_1 g_1 & \mathcal{G}_0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ D_K g_K & 0 & \ldots & \mathcal{G}_0 \end{bmatrix} \tag{9}$$

and

$$\begin{aligned} x_{ap} &= [x_0^{(0)}, x_1^{(1)}, ..., x_K^{(1)}]^T \\ \mathcal{B}_{ap} &= [\mathcal{B}, 0, ..., 0]^T \\ \mathcal{L}_{ap} &= [\mathcal{B}, \delta D_1\mathcal{B}, ..., \delta D_K\mathcal{B}]^T. \end{aligned}$$

Note that $\mathcal{C}_{ap}$ has the same lower-triangular structure as $\mathcal{G}_{ap}$ does.

The state variable y_{ap} at output for those critical tiles can be also divided into two parts:

$$y^{(0)} = y_0^{(0)} \ (\in R^1), \quad y^{(1)} = \{y_1^{(1)}, ..., y_K^{(1)}\} \ (\in R^K).$$

As a result, solving (8) results in the nominal value, $y^{(0)}$, and its first-order sensitivity $y^{(1)}$ with respect to each parameter D_i.

The large system equation (8) can be reduced using projection with preserved moments (of s) up to the qth-order. Because $\mathcal{G}_{ap}$ and $\mathcal{C}_{ap}$ have lower-triangular structures, a flat projection matrix Q_{ap} can be constructed recursively using an iterative Arnoldi method [19]. However, [19] directly projects (8) by Q_{ap}. It leads to a reduced macromodel losing the block structure for both state matrices and variables. As a result, $y^{(0)}$ and $y^{(1)}$ are interleaved with each other.

In this paper, instead of using the flat projection matrix Q_{ap}

$$Q_{ap} = [Q_0, \underbrace{Q_1, ..., Q_K}_{K}], \tag{10}$$

we introduce a structured projection matrix

$$\mathcal{Q}_{ap} = \begin{bmatrix} Q_0 & 0 & \cdots & 0 \\ 0 & Q_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_K \end{bmatrix} \tag{11}$$

by partitioning Q_{ap} according to the dimension of $x^{(0)}$ and $x^{(1)}$. Note that $\mathcal{G}_{ap}$, $\mathcal{C}_{ap}$ and $\mathcal{Q}_{ap}$ have larger dimensions than G, C and Q. However, as $\mathcal{G}_{ap}$, $\mathcal{C}_{ap}$ and $\mathcal{Q}_{ap}$ are in block triangular or block diagonal form, they can be implemented in a block-matrix data structure [20, 21] without memory usage increase.

More importantly, as shown in [20, 21], the projection by $\mathcal{Q}_{ap}$ preserves the block matrix structure. As a result, the order-reduced state matrices after projection by $\mathcal{Q}_{ap}$ are

$$\widetilde{\mathcal{G}}_{ap} = \mathcal{Q}_{ap}^T\mathcal{G}_{ap}\mathcal{Q}_{ap} \quad \widetilde{\mathcal{C}}_{ap} = \mathcal{Q}_{ap}^T\mathcal{C}_{ap}\mathcal{Q}_{ap}$$
$$\widetilde{\mathcal{B}}_{ap} = \mathcal{Q}_{ap}^T\mathcal{B}_{ap} \quad \widetilde{\mathcal{L}}_{ap} = \mathcal{L}_{ap}\mathcal{Q}_{ap}$$

where

$$\widetilde{\mathcal{G}}_{ap} = \begin{bmatrix} Q_0^T\mathcal{G}_0Q_0 & 0 & \ldots & 0 \\ D_1Q_1^Tg_1Q_0 & Q_1^T\mathcal{G}_0Q_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ D_KQ_K^Tg_KQ_0 & 0 & \ldots & Q_K^T\mathcal{G}_0Q_K \end{bmatrix} \tag{12}$$

Note that the reduced $\widetilde{\mathcal{C}}_{ap}$ has the same preserved lower-triangular structure as $\widetilde{\mathcal{G}}_{ap}$.

In addition, the structured and parameterized macromodel

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

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 $Q_{ap} \subseteq \mathcal{Q}_{ap}$, a q-th ordered projection by $\mathcal{Q}_{ap}$ still preserves at least q moments according to [17].

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

$$(\widetilde{\mathcal{G}}_{ap} + \frac{1}{h}\widetilde{\mathcal{C}}_{ap})\widetilde{x}_{ap}(t) = \frac{1}{h}\widetilde{\mathcal{C}}_{ap}\widetilde{x}_{ap}(t-h) + \widetilde{\mathcal{B}}_{ap}u(t)$$
$$\widetilde{y}_{ap}(t) = \widetilde{\mathcal{L}}_{ap}^T\widetilde{x}_{ap}(t). \tag{13}$$

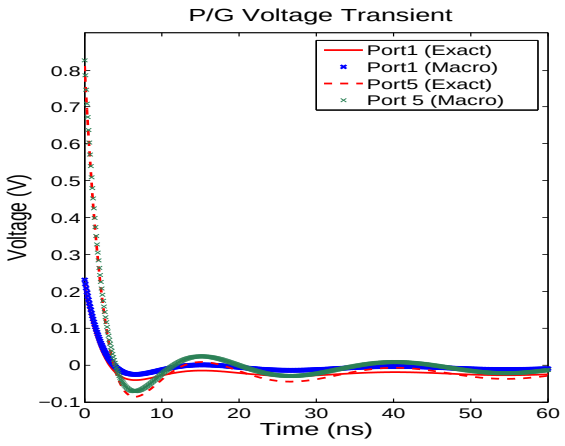

Figure 6: Transient voltage responses of exact models and our macromodels(order=40) at port 1 and 5 of layer-1. The macromodels are visually identical to those exact models.

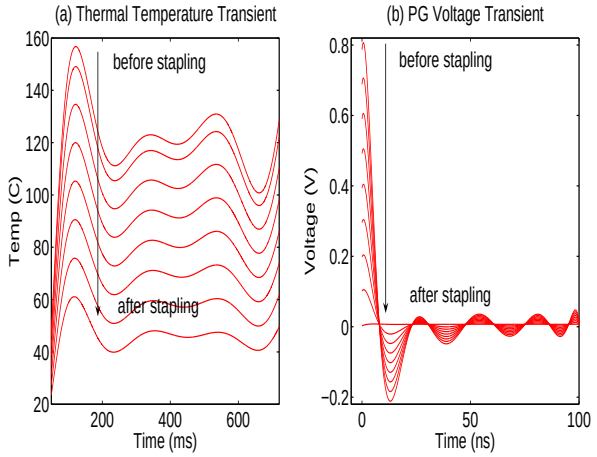

Figure 7: Iterative optimizations showing the reduction of (a) temperature and (b) voltage violation by via-stapling.

where

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

The overall voltage/temperature (V/T) vector at those critical tiles perturbed by the via-density vector, i.e. $\mathbf{D}$, is

$$\widetilde{y}(\mathbf{D}, t) = \widetilde{y}^{(0)}(\mathbf{D}, t) + \widetilde{y}^{(1)}(\mathbf{D}, t). \quad (14)$$

Since our reduction preserves the block structure, the reduced nominal value $\widetilde{y}^{(0)}$ and first-order sensitivity $\widetilde{y}^{(1)}$ at output (critical tiles) can be solved independently. Moreover, because the reduced system still has the lower-triangular structure, it is obvious that (13) can be efficiently solved using back substitution with only one LU factorization of

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

As a result, such a structured and parameterized macromodel can be incorporated in a sensitivity based optimization for efficient yet accurate staple vias.

4. SENSITIVITY BASED OPTIMIZATION

Because our structured and parametrized macromodel provides both nominal values and sensitivities, they can be incorporated in any gradient-based optimization. However, the Hessian matrix used in gradient-based optimization [22] is computationally expensive to obtain second-order gradients. If there are K parameters in the design, it needs K^2 parameterized second-order moments to generate a Hessian matrix. As a result, the cost to build and simulate a macromodel becomes huge. This is inefficient and not necessary for the via-planning during the early-design stage.

For the sake of speed to handle large scale problems, the technique used in our paper is a sensitivity-based heuristic similar to the TILOS [23] algorithm. By successively increasing via density for the via-level with the largest gain in each iteration, our algorithm staples a minimum number of vias to reduce both temperature and voltage violations in problem formulation (2). This greedy heuristic optimizer can solve large scale designs efficiently and effectively.

The overall optimization to solve problem formulation (2) is outlined in Algorithm 1. Note that the weighted sensitivity vector $\mathbf{S}$ is a weighted-sum of normalized voltage-sensitivity vector $\mathbf{S}_V$ and thermal-sensitivity $\mathbf{S}_T$:

$$\mathbf{S} = \alpha \cdot \mathbf{S}_T/||\mathbf{S}_T|| + \beta \cdot \mathbf{S}_V/||\mathbf{S}_V||. \quad (15)$$

where α and β are weights for $\mathbf{S}_T$ and $\mathbf{S}_V$. The via density vector $\mathbf{D}$ is updated by

$$\mathbf{D}^{(iter+1)} = \mathbf{D}^{(iter)} + \gamma^{(iter)} \cdot \mathbf{S}^{(iter)},$$

where γ is an adaptive-controlled step size and decreases as the iteration proceeds.

Algorithm 1 Via Stapling Procedure

Input: Critical nodes, via pattern number K, signal congestion induced bound D_{max} and current-density induced bound D_{min}
Step 0: Reduce (8) using structured and parameterized reduction;
Step 1: Compute nominal voltage(V)/temperature(T) and sensitivity $\mathbf{S}_V/\mathbf{S}_T$ using (13) by backward-substitution;
Step 2: Determine V_{max} and T_{max} of critical nodes;
Step 3: Increase the via density $\mathbf{D}$ according to weighted sensitivity $\mathbf{S}$ in the range of (D_{min}, D_{max});
Step 4: Update the structured and parameterized macromodel according to (12);
Step 5: Repeat from Step 1 until maximum noise/temperature constraints are satisfied;
Output: Via density vector $\mathbf{D}$

Because our macromodel is parameterized, only one reduction is needed and the reduced state matrices can be repeatedly used when updating the parameter vector $\mathbf{D}$ according to (12). In addition, the reduced model is much smaller than the original one and has a lower-triangular structure. Its nominal value and sensitivities, therefore, can be efficiently solved by backward-substitution of (13) with only one LU factorization. As a result, the optimization procedure in Algorithm 1 is computationally efficient.

5. EXPERIMENTS

The proposed modeling and algorithm have been implemented in C. Experiments are run on a Sun-Fire-V250 workstation with 2G RAM, and the reported number of vias are all for no-signal vias. Silicon, copper and dielectric are assumed for via, heatsink, active device layer, inter-layer and PG plane, respectively. Table 3 and Table 4 summarize their electrical and thermal constants and dimensions. In addition, 20% of the deive-layer tiles have a random input of thermal-power in the range of 1 to $5 \times 10^6 W/m^2$. Their clock gating pattern has a period of 100ms where the power in the standby mode is 20% of the running mode. 10% of power/ground-plane tiles have random inputs of switching-currents in the range of 1 to $5 \times 10^{-1} A$ with rising-time $0.01ns$. All power sources are uniformly distributed across the active device layer or PG-plane. The range of via density is set from 100 to 80000 for each level, and the weight α and

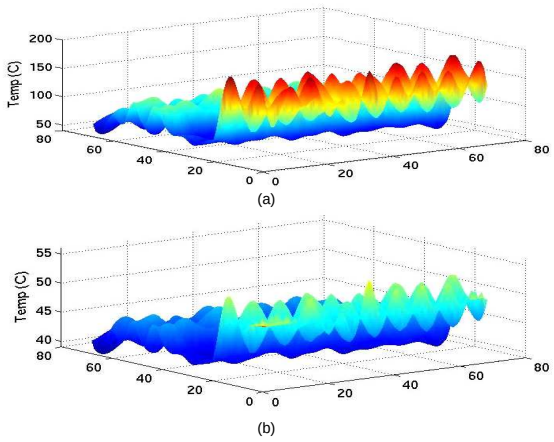

Figure 8: Steady-state temperature map of bottom device layer (a) before via-stapling, and (b) after via-stapling in a different temperature scales.

	Silicon	Copper	Dielectric
σ	NA	$59.6 \times 10^6 S/m$	NA
ϵ_r	NA	NA	3.3
μ_r	NA	NA	1.0
κ_R	$100W/m \cdot K$	$400W/m \cdot K$	$50W/m \cdot K$
κ_C	$1.75 \times 10^6 J/m^3 \cdot K$	$3.55 \times 10^6 J/m^3 \cdot K$	$0.7W/m \cdot K$

Table 3: Electrical and thermal constants.

β are equal for (15). A modest 3D stacking with 1-heat-sink, 2-device-layer, 2-dielectric-layer and 2-P/G-plane is assumed for the experiments.

5.1 Verification of Macromodel and Optimization

The first example is a 3D stacking with 7900 total tiles and a level-4 via pattern. Fig. 6 compares the time-domain waveform between our macromodel and exact MNA solution at ports 1 and 5 of one selected P/G plane. The macromodel (expanded at $s_0 = 0$) uses order 40. Clearly, our macromodel is accurate.

Fig. 7 presents the decreasing of the transient temperature and voltage violation during the simultaneous optimization procedure. Fig. 8 further shows the steady-state temperature map across the top device layer. In this example, we assume that all thermal-power sources are located at one side of the device layer. The initial chip temperature at the top layer is around $150°C$, and its temperature profile at steady-state is shown in Fig. 8 (a). In contrast, the via-stapling results in a cooler temperature that closely approaches the targeted temperature as shown in Fig. 8 (b).

5.2 Comparison between Steady-state and Transient Thermal Analysis

We further compare the runtime and the number of vias between the steady-state and transient thermal analysis. We increase the circuit complexity by increasing the number of discretized tiles, and need more levels for vias when the tile number becomes larger. The sequential optimization in this comparison is used. We first allocate vias to satisfy the power integrity constraints with targeted voltage violation V_{max} of $0.2V$, and then allocate vias to satisfy the thermal integrity constraints with targeted temperature T_{max} of $52°C$ and considering the already stapled "power/ground" vias for heat removal.

layer	size	material
heat-sink	$2cm \times 2cm \times 1mm$	copper
device-layer	$1cm \times 1cm \times 4um$	silicon
inter-layer	$1cm \times 1cm \times 1um$	dielectric
P/G plane	$2cm \times 2cm \times 10um$	copper

Table 4: Dimensions of 3D ICs layers.

Table 5 presents the results. The vias are over-designed when using steady-state analysis. Compared to the optimization by steady-state analysis, the optimization by transient thermal analysis reduces vias by 11.5% for a circuit with 27740 tiles. This is because the steady-state analysis has to assume a constant maximum thermal-power input for all tiles in order to get a valid solution. In reality, it seldom happens that all tiles have their maximum input at the the same time. In contrast, our transient thermal analysis can accurately generate the transient temperature using the input of transient thermal-power.

Furthermore, directly solving state equations as in [12, 13] results in longer runtimes. In contrast, the macromodel can efficiently match the transient response using around 20 moments. For the same circuit with 27740 tiles, our macromodel has a 155X smaller runtime compared to the steady-state analysis, and the steady-state analysis can not complete the largest example.

5.3 Comparison between Sequential and Simultaneous Optimizations

We further compare the sequential optimization with the simultaneous thermal/power optimization, and first discuss via patterns for thermal and power integrity, respectively. As shown in Table 6, for a circuit with 27740 tiles, when only using the thermal-constraint, more vias tend to be stapled for high-level patterns. As a higher level pattern means more uniform via distribution, the thermal constraint results in a more uniformly distributed via pattern. On the other hand, when only using P/G-constraint, more vias are stapled in the center, i.e., using level-0 via pattern to reduce the power/ground plane loop inductance or SSN. Due to such an opposite stapling trend, a via-stapling in a sequential fashion results in excessive number of vias. In contrast, the vias are distributed more uniformly in all levels when simultaneously considering the thermal and power integrity. Finally, we compare the results using simultaneous optimization and sequential optimization. On average, our simultaneous optimization further reduces 34% vias compared to the the sequential optimization by steady-state analysis in Table 5, and reduces 22.5% vias compared to the sequential optimization with transient analysis.

6. CONCLUSIONS

Inter-layer non-signal vias are effective to reduce power supply noise and temperature hotspots in 3D ICs. Existing work on via-stapling [6–8, 12, 13] does not consider thermal and power integrity simultaneously, and uses steady-state thermal analysis. To reduce the number of vias for targeted power and thermal integrity, this paper has presented the first in-depth study on simultaneous power and thermal integrity driven via-stapling.

Our simultaneous power and thermal integrity optimization minimizes non-signal vias subject to constraints on transient temperature and voltage violations, which are calculated by a structured and parameterized model reduction. This model reduction also generates parameterized sensitivities of temperature and voltage violations with respect to the via pattern and density. The resulting macromodel is used in the efficient greedy optimization simultaneously driven by power and thermal integrity.

Experiments with two active device layers show that for the sequential power and thermal optimization, using the transient thermal analysis reduces non-signal vias by on average 11.5% compared to using the steady-state thermal analysis. In addition, our simultaneous optimization of power and thermal integrity reduces on average 34% non-signal vias compared to the existing approach assuming the sequential optimization and steady-state thermal analysis. The via reduction could be higher for the 3D design with more device layers.

The power integrity in this paper considers noise on power/ground planes in the package without considering on-chip power supply routing. In the future, we will develop simultaneous power and signal routing to optimize on-chip power and thermal integrity.

total tile#	level vector	Steady-state(direct)		Transient(MACRO)			
		solve dc (s)	total via # by seq-opt	redu ckt (s)	solve BE(s)	total via # by seq-opt	total via # by sim-opt
620	0,1	4.06	176877	0.01	0.12	156154 (-11%)	118020 (-32%)
2140	0,1,2	26.37	187422	0.13	0.17	166971 (-11%)	127651 (-32%)
7900	0,1,2,3	167.9	235484	1.22	0.86	206482 (-12%)	140433 (-36%)
27740	0,1,2,3,4	1243.7	239379	5.12	1.07	211184 (-12%)	143718 (-37%)
55680	0,1,2,3,4,5	NA	NA	15.87	3.65	216732 (NA)	144998 (NA)
						avg. (-11.5%)	avg. (-34%)

Table 5: Comparisons of via number and runtime for the sequential optimization with steady-state analysis, the sequential optimization with transient analysis and the simultaneous optimization with transient analysis.

opt-method	level				
	0	1	2	3	4
P/G-only	76832	3410	1901	876	/
Thermal-only	/	1157	43567	4007	79432
Simultaneous	67058	811	2500	2808	70541

Table 6: Comparisons of via distribution at different levels.

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] J. U. Knickerbocker and et. al., "Development of next-generation system-on-package technology based on silicon carriers with fine-pitch chip interconnection," *IBM Research Journal (Power and Packaging)*, pp. 725–754, 2005.

[4] M. Swaminathan, J. Kim, I. Novak, and J. Libous, "Power distribution networks for System-on-Package: Status and challenges," *IEEE Trans. on Advanced Packaging*, pp. 286–300, 2004.

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

[6] J. Zhao, J. Zhang, and J. Fang, "Effects of power/ground via distribution on the power/ground performance of C4/BGA packages," in *Proc. IEEE Electrical Performance of Electronic Packaging (EPEP)*, 1998.

[7] Y. Hong and et. al., "Analysis for complex power distribution networks considering densely populated vias," in *Proc. Int. Symp. on Quality Electronics and Design (ISQPED)*, 2005.

[8] C. Ryu and et. al., "High frequency electrical circuit model of chip-to-chip vertical via interconnection for 3D chip stacking package," in *Proc. IEEE Electrical Performance of Electronic Packaging (EPEP)*, 2005.

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

[10] 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 *Proc. Int. Symp. on Low Power Electronics and Design (ISLPED)*, 2005.

[11] 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.

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

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

[14] A. E. Ruehli, "Equivalent circuits models for three dimensional multiconductor systems," *IEEE Trans. on Microwave Theory and Techniques*, pp. 216–220, 1974.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] H. Yu, L. He, and S. Tan, "Block structure preserving model reduction," in *IEEE International Workshop on Behavioral Modeling and Simulation (BMAS)*, 2005.

[21] H. Yu, Y. Shi, and L. He, "Fast analysis of structured power grid by triangularization based structure preserving model order reduction," in *Proc. Design Automation Conf. (DAC)*, 2006.

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

[23] J. P. Fishburn and A. E. Dunlop, "TILOS: A posynomial programming approach to transistor sizing," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 326–328, 1985.