Allocating power ground vias in 3D ICs for simultaneous power and thermal integrity

Yu, Hao; Ho, Joanna; He, Lei

2009

Yu, H., Ho, J., & He, L. (2009). Allocating power ground vias in 3D ICs for simultaneous power and thermal integrity. *ACM Transactions on Design Automation of Electronic Systems*, 14(3).

<https://hdl.handle.net/10356/94215>

<https://doi.org/10.1145/1529255.1529263>

© 2009 ACM. This is the author created version of a work that has been peer reviewed and accepted for publication by *ACM Transactions on Design Automation of Electronic Systems*, ACM. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [<http://dx.doi.org/10.1145/1529255.1529263>].

Downloaded on 26 Apr 2023 18:09:31 SGT

Allocating Power Ground Vias in 3D ICs for Simultaneous Power and Thermal Integrity

HAO YU

Berkeley Design Automation

and

JOANNA HO and LEI HE

University of California, Los Angeles

The existing work on via allocation in 3D ICs ignores power/ground vias' ability to simultaneously reduce voltage bounce and remove heat. This paper develops the first in-depth study on the allocation of power/ground vias in 3D ICs with simultaneous consideration of power and thermal integrity. By identifying principal ports and parameters, effective electrical and thermal macro-models are employed to provide dynamic power and thermal integrity as well as sensitivity with respect to via density. With the use of sensitivity, an efficient via allocation simultaneously driven by power and thermal integrity is developed. Experiments show that compared to sequential power and thermal optimization using static integrity, sequential optimization using the dynamic integrity reduces non-signal vias by up to 18%, and simultaneous optimization using dynamic integrity further reduces non-signal vias by up to 45.5%.

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

Additional Key Words and Phrases: Thermal and Power Integrity, Macromodeling, Parametric 3D-IC Design

1. INTRODUCTION

In today's two dimensional (2D) Systems-on-Chip (SoC) integration, the memory arrays are surrounded by logic circuits to address bits and to perform logic functions. Its performance thereby is limited by the long interconnect length. Thanks to advances in recent process technology, the introduction of the three dimensional (3D) integration [Banerjee et al. 2001; Goplen and Sapatnekar 2003; Das 2004; Cong et al. 2004; Davis and et al. 2005; Knickerbocker and et. al. 2005; Lim 2005; Goplen and Sapatnekar 2005; Cong and Zhang. 2005; Sapatnekar 2006; Li et al. 2006; Yu et al. 2006; Yu et al. 2006] enables flash, DRAM and SRAM to be placed atop logic devices and even microprocessor cores in SoCs. Compared

Author's address: Hao Yu is now with Berkeley Design Automation, Santa Clara, CA 95054 USA (e-mail: hao.yu@berkeley-da.com). His work was performed at University of California, Los Angeles. Joanna Ho are Lei He are with Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: lhe@ee.ucla.edu).

Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.

© 2009 ACM 1084-4309/2009/0400-0001 \$5.00

to the 2D integration, such an upwards arrangement of interconnect in 3D results in not only a shorter communication length but also a denser device density for tomorrow's SoCs. Apparently, the physical (or geometrical) arrangement of layout resources in 3D would be similar to the physical design in 2D. However, as there are large numbers of devices densely packed in a number of device layers, it brings a significant burden to remove the heat and deliver the power (supply voltage) in 3D ICs.

Fig. 1 illustrates a typical 3D stacking of multiple device layers within one package. The supply voltage is delivered from the bottom power/ground planes in the package, passed by the through vias and C4 bumps and connected to the on-chip power/ground grid on active device layers. We call the through vias that deliver the supply voltage as the *power/ground vias*. The 3D integration, by definition, has integrated more than one layer of the active device. They draw a much larger current from the package power/ground planes than the 2D ICs. This obviously can result in the IR drop for the horizontal on-chip power/ground grid. The surge of the injecting current can further lead to a large simultaneous switching noise (SSN) for those I/O drivers at the chip package interface. Fig. 2 shows a detailed view of how to place signal and power/ground vias through package planes. Clearly, the region to place power vias or ground vias decides the path of the returned current and the loop area for those signal nets connecting I/Os. As such, they form various loop-inductances with different sizes that have significant couplings with each other. Moreover, when the region to place power/ground vias is determined, the number of placed vias in that location, i.e., density, decides the amount of coupled current. When there are more vias in one region, its current density is lower and its self-inductance is smaller. This leads to a smaller SSN at I/Os. We call the voltage bounce at I/Os *power integrity* in this paper. To satisfy the power-integrity constraints for I/Os, the allocation of power/ground vias have been studied in [Zhao et al. 1998; Hong and et. al. 2005; Ryu and et. al. 2005]. Compared to the allocation of off-chip decoupling capacitors (mm^2), placing vias (um^2) have a smaller cost of area. This is important for high-density integration in 3D ICs as there may be a large number of signal nets to be delivered from the package to the chip. Therefore, a thorough study of power/ground vias is needed for 3D ICs.

Due to the increased power density and the slow heat-convection at inter-layer dielectrics, the heat dissipation is another concern in 3D ICs [Banerjee et al. 2001]. The excessively high temperature can significantly degrade the reliability and performance of interconnects and devices [Teng et al. 1997; Chiang et al. 2001; Wang and Chen 2003; Li et al. 2004; Zhan and Sapatnekar 2007]. We call the temperature gradient at active device layers *thermal integrity* in this paper. As shown in Fig. 1, a heat-sink is placed on the top of device layers and it is the primary heat-removal path to the ambient air. Note that through vias deliver supply voltages or signals from the bottom package through each active device layer. Since metal vias are good thermal conductors, the through vias can provide additional heat-removal paths passing the inter-layer dielectrics to the top heat-sink. This leads to the concept of adding *dummy thermal vias* or *thermal vias* directly inside chips [Chiang et al. 2001] to reduce effective thermal resistances. Its physical arrangement has been studied further [Goplen and Sapatnekar 2003; Cong et al. 2004; Goplen and

Sapatnekar 2005; Cong and Zhang. 2005; Li et al. 2006; Yu et al. 2006; Yu et al. 2006].

The dummy vias namely need to occupy the routing resources for signal nets and power/ground nets. Note that the use of power/ground vias is a must to deliver supply voltage for those active device layers from the package. They also provide good thermal conduction paths when extended to connect the heat-sink. This motivates us to reuse the power/ground vias for the heat removal in 3D ICs. The reusing of power/ground vias as thermal vias would avoid an over-design that separately allocates power/ground vias and thermal vias. Though there could be other known design freedoms such as decaps, a need is required to study the optimization of power and of thermal integrity simultaneously because power/ground vias happen to be useful to optimize both power and thermal integrity in 3D ICs.

As such, an allocation of power/ground vias simultaneously driven by power and thermal integrity can be defined for 3D ICs. Similar to [Goplen and Sapatnekar 2005], the allocation of power/ground vias in this paper is assumed after the placement and global routing and before the detailed routing for signal nets. The active device layers and power/ground layers are uniformly discretized into a number of tiles. The *input ports* are defined for those tiles where to inject input signals. The *output ports* are defined for those tiles where to probe the integrity constraint. The output ports are similar to the pre-designate regions in [Goplen and Sapatnekar 2005], all aligned for adding and adjusting vias. The *tracks* are defined accordingly for the vertical space occupied by the vertical through vias going from the bottom package to the top heat-sink. The number of tracks for power/ground vias, i.e., the via density, is determined by both integrity constraints as well as the signal-net congestion in the same track.

However, simultaneously considering power and thermal integrity is obviously a challenging task. Firstly, in the modern VLSI designs, dynamic power management such as clock-gating and uncertainty from the workload can lead to time-varying power inputs. This results in a spatially and temporally variant thermal model. The temperature gradient can have either a sharp-transition with a large peak value, or a time-accumulated impact to the device reliability. In addition, different regions can reach their worst-case temperature at different times. A dynamic thermal-integrity constraint thereby is needed to accurately guide the physical level resource allocation. A dynamic thermal model with the time-varying thermal power input is studied in [Teng et al. 1997; Chiang et al. 2001; Wang and Chen 2003; Li et al. 2004; Yu et al. 2006; Yu et al. 2006; Zhan and Sapatnekar 2007]. Compared to the use of the steady-state thermal model [Goplen and Sapatnekar 2003; Cong et al. 2004; Goplen and Sapatnekar 2005; Cong and Zhang. 2005; Li et al. 2006], the dynamic thermal model not only accurately monitors the temperature gradient but also avoids the over-design caused by the pessimistic static thermal-integrity constraint. Note that the dynamic power-integrity has already been employed in many on-chip or off-chip power integrity verifications and designs [Zhao et al. 2002; Zhao et al. 2002; Su et al. 2003; Li and et. al. 2005; Chen and He 2006].

Next, an accurate electrical or thermal model considering all device layers and package planes can result in millions of unknowns with thousands of time-varying input sources. Moreover, during the physical design, the design parameters such

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

as the device size (or density) and location further increase the complexity. This makes the design for power and thermal integrity difficult to accomplish in a reasonable time and hence further blocks the physical level optimization. To reduce the simulation cost, the macromodel based approaches [Zhao et al. 2002; Su et al. 2003; Zheng et al. 2003; Chen and He 2006; Yu et al. 2006; Yu et al. 2006] are used to find a compact representation of the electrical and thermal model. The works in [Yu et al. 2006; Yu et al. 2006] further introduce a structured and parameterized macromodel to provide both the nominal response and the sensitivity for the purpose of optimization.

The macromodeling based allocations in [Yu et al. 2006; Yu et al. 2006], however, are not effective when the the number of input ports or output ports is large. Generally, there can be thousands of thermal-power sources injected at each active layers or hundreds of switching-current sources injected I/Os. The size of the macromodel increases with the number of ports, and hence the computational cost to solve the macromodel is still high. Moreover, the effectiveness to apply the macromodel also heavily depends on how to scope with the design parameters. Here the parameters are the via densities at tracks. Assuming that the locations of tracks are pre-designated [Goplen and Sapatnekar 2005], [Yu et al. 2006] adjusts the via density in each track. [Yu et al. 2006] further decomposes those tracks into multiple groups described by levels. The vias are then allocated uniformly for those tracks in the same level. However, it is still expensive to probe the integrity and adjust the via density at each track for a large number of pre-designated tracks.

In this paper, we introduce an allocation of power/ground vias to simultaneously consider the dynamic power and thermal integrity in 3D ICs. The primary contributions are two-fold. Firstly, we notice that the previous thermal via allo-

Fig. 2. The power delivery by vertical power/ground vias and its impact to inductive current loops.

cations [Cong et al. 2004; Goplen and Sapatnekar 2005; Cong and Zhang. 2005; Li et al. 2006; Yu et al. 2006] assume adding dummy vias to conduct heat. They ignore the fact that power/ground vias can help heat-removal as well. Therefore, the reusing of the power/ground via as thermal via can save the routing resource for signal nets. In this paper, we present a design methodology that reuses the power/ground vias as the heat-removal paths. The allocation of power/ground vias are oriented to minimize not only the dynamic power integrity, i.e., the voltage bounce for those I/Os at package and chip interface, but also the thermal integrity, i.e., the temperature gradient at those active device layers.

Secondly, we develop an effectively parameterized macromodel. The state variables such as voltages and temperatures, the input ports and output ports, and the parameters are all compressed by their dominant subspace. By taking ‘snapshots’ of input and output ports, we extract the correlations for input and output ports, respectively. The subspaces are then extracted from the correlations and are applied to compress input and output ports. Moreover, recall that those outputs are the places to probe integrity and add vias. We introduce a spectral clustering to further compress the parameters by the extracted subspace from the output port correlation. As a result, we build a system with a small number of principal ports, and parametrize it by a small number of principal parameters. By further employing a structure and parameterized macromodel, the voltages, temperatures and their sensitivities are further compressed. The macromodel is embedded in the sensitivity based allocation, where the via density is adjusted according to the sensitivities. Compared to the works in [Yu et al. 2006; Yu et al. 2006], the method developed in this paper improves the effectiveness of the macromodeling by identifying the principal ports and parameters.

The use of dummy vias in [Cong et al. 2004; Goplen and Sapatnekar 2005; Cong and Zhang. 2005; Li et al. 2006; Yu et al. 2006] separates the allocation of power/ground vias from the thermal vias. We call the separated allocation of thermal vias and power/ground vias as the *sequential optimization*, and call our allocation of power/ground vias for both power and thermal integrity as the *simultaneous optimization*. Moreover, the steady-state analysis is employed to calculate

a static integrity [Cong et al. 2004; Goplen and Sapatnekar 2005; Cong and Zhang. 2005; Li et al. 2006]. We use the sequential optimization with the static integrity as the baseline, in comparison to the sequential optimization with the dynamic integrity [Yu et al. 2006] and the simultaneous optimization with the dynamic integrity proposed in this paper. As shown by experiments, for the sequential integrity optimization, the use of dynamic integrity reduces up to 18% allocated non-signal vias compared to the use of static integrity. Furthermore, our simultaneous optimization using dynamic integrity reduces up to 44.5% non-signal vias compared to the baseline.

The rest of the paper is organized as follows. We review the background of the dynamic electrical and thermal model and their macromodel by model order reduction in Section 2. We formulate the allocation problem that simultaneously considers the power and thermal integrity in Section 3. We introduce the principal ports characterization in Section 4, and present the via-allocation algorithm using macromodels in Section 5. We present experiment results in Section 6 and conclude in Section 7.

2. BACKGROUND OF ELECTRO-THERMAL MODEL

2.1 Circuit Model for Dynamic Integrity

Since the analysis of thermal integrity is targeted during physical design, the granularity of discretized thermal model is thereby smaller than those for the micro-architecture design [Liao et al. 2005]. As such, a distributed thermal- RC circuit is employed for active device layers and dielectric layers. Each silicon or dielectric layer is uniformly discretized by the finite difference method. As shown in Fig. 3 (a) and (b), each discretized tile can be represented by an RC -cell. The thermal vias and global signal nets are modeled by the discretized RC as well. The following boundary condition is added. The packaging (C4-bump package) and heat-sink are modeled by a simple 1D convection resistor connected to the thermal ground, i.e., the ambient air with a fixed temperature (See Fig. 3 (d)). In addition, the inputs are the time-varying thermal power [Tiwari et al. 1998; Liao et al. 2005] 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), and injected at input ports of each layer. As such, a temporally and spatially variant temperature at output ports can be considered by defining an *integrity integral* with respect to time and space [Yu et al. 2006].

Since the active device layer at the bottom (See. Fig. 1) has the longest path to the heat-sink on the top, in this paper, a dynamic thermal integrity is defined as the integrated temperature fluctuation at p_o output ports on the bottom device layer. As shown in [Teng et al. 1997; Chiang et al. 2001; Wang and Chen 2003; Li et al. 2004; Yu et al. 2006; Yu et al. 2006; Zhan and Sapatnekar 2007], a dynamic thermal integrity can accurately capture not only the sharp-transition of temperature change due to the dynamic power management, but also the time-accumulated temperature impact that can affect the device reliability. Similar to the static thermal-integrity analysis, the dynamic thermal integrity assumes the worst-case input from a limited number of thermal-power inputs. However, since the dynamic integrity has a more accurate transient temperature profile, it leads to a smaller allocation compared to

Fig. 3. (a) has 2-tile and 1-via, (b) is its equivalent electrical *RLC*-cell, (c) is its equivalent thermal *RC*-cell, and (d) is the entire distributed thermal *RC* model.

the static thermal-integrity based design [Goplen and Sapatnekar 2003; Cong et al. 2004; Goplen and Sapatnekar 2005; Cong and Zhang. 2005; Li et al. 2006].

Moreover, the dynamic power integrity [Zhao et al. 2002; Su et al. 2003; Li and et. al. 2005; Chen and He 2006; Zhao and et.al. 2007] in this paper is defined as the time-integrated voltage bounce at power/ground I/Os, which are located on the interface between the bottom device layer and the package. The discretized tile of a power/ground plane, power/ground via, and signal via in the package are modeled by an *RLC*-cell (See Fig. 3 (a) and (c)) under the PEEC model [Ruehli 1974]. The C4-bump, on-chip global signal nets and power grid are modeled as *RC* circuits. The inputs are the switching-current for those input signals in the package. In addition, the power/ground vias in this paper are designed to provide power supply, alter voltage bounce, and remove heat from the package planes to the heat-sink. To avoid the difference in I/R drop, the vertical power/ground vias are aligned along the vertical dimension. As a result, the power/ground vias reused for heat-removal are all aligned from the bottom package to the top heat-sink.

The extracted thermal-*RC* and electrical-*RLC* models can be described in the state equation under modified nodal analysis (MNA). Since the distributed thermal-*RC* circuit is a simpler case for the electrical-*RLC* circuit, we use the presentation of the *RLC* circuit unless stated otherwise. The MNA for the nominal electrical-*RLC* circuit in time-domain without power/ground vias is

$$\begin{aligned} \mathcal{G}x(t) + \mathcal{C} \frac{dx(t)}{dt} &= \mathcal{B}I(t) \\ y(t) &= \mathcal{L}^T x(t) \end{aligned} \quad (1)$$

N	Dimension of MNA state equation
N_v	Dimension of voltage variables in MNA
N_l	Dimension of current variables in MNA
p_i	Dimension of input-ports
p_o	Dimension of output-ports
$x(y) (\in R^{N \times 1})$	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
$\mathbf{E}_i (\in R^{N_v \times p_i})$	Input/output incident matrix
$\mathbf{E}_l (\in R^{N_v \times N_l})$	Inductive incident matrices
$X_i (\in R^{N \times N})$	Parameterized topological matrix for

Table I. Notation list for system equation (2). Note that $N = N_v + N_l$.

or in frequency (s) domain

$$\begin{aligned} [\mathcal{G} + s\mathcal{C}]x(s) &= \mathcal{B}\mathbf{I}(s) \\ y(s) &= \mathcal{L}^T x(s) \end{aligned} \quad (2)$$

where

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

and

$$\mathcal{G} = \begin{bmatrix} G & \mathbf{E}_l \\ -\mathbf{E}_l^T & 0 \end{bmatrix}, \mathcal{C} = \begin{bmatrix} C & 0 \\ 0 & L \end{bmatrix}. \quad (3)$$

Note that $\mathcal{B}$ is the topology matrix to describe p_i input ports with injected input sources, and $\mathcal{L}$ is the one to describe p_o output ports for probing integrity and adjusting via density. All notations in (3) are summarized in Table 1.

Differences exist between thermal- RC and electrical- RLC circuits in MNA. For a thermal- RC circuit, (3) becomes

$$\mathcal{G} = G, \quad \mathcal{C} = C \quad (4)$$

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

2.2 Macromodeling by Model Order Reduction

As the layouts of active device layers and power/ground planes are discretized sufficiently, the size of resulting state-matrices thereby can be huge. As such, the distributed thermal- RC and electrical- RLC model are difficult to be employed for either the integrity verification or the optimization.

Model order reduction [E.J.Grimme 1997; Odabasioglu et al. 1998] finds the dominant state variables and obtains compact macromodels. As shown in [E.J.Grimme

1997; Odabasioglu et al. 1998], the dominant state variables are related to the block Krylov subspace

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

constructed from moment matrices

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

by expanding the system transfer function

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

at one frequency s_0 .

By applying the block Arnoldi iteration [Odabasioglu et al. 1998], a small dimensioned projection matrix Q ($N \times q \times p_i$) can be found to contain q th-order block Krylov subspace

$$\mathcal{K}(\mathcal{A}, \mathcal{R}, q) \subseteq Q.$$

Using Q the original system can be reduced by projection

$$\hat{\mathcal{G}} = Q^T\mathcal{G}Q \quad \hat{\mathcal{C}} = Q^T\mathcal{C}Q \quad \hat{\mathcal{B}} = Q^T\mathcal{B} \quad \hat{\mathcal{L}} = Q^T\mathcal{L}.$$

Accordingly, the reduced system transfer function becomes

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

As proven in [E.J.Grimme 1997; Odabasioglu et al. 1998], the reduced $\hat{H}$ approximates the original system transfer function $H(s)$ by matching the first q block moments expanded at the frequency-point s_0 . This procedure can be applied to generate compact macromodels for both thermal-*RC* and electrical-*RLC* circuits.

Our thermal-*RC* and electrical *RLC* models have multiple inputs and multiple outputs (MIMO). This brings challenges for the projection based model order reduction. The dimension of the reduced MIMO system $\hat{H}$ ($\in R^{p_o \times p_i}$) depends on the input-port number p_i and p_o . In general, there are large numbers of injecting inputs. An accurate monitoring of integrity also needs large numbers of pre-designate regions to probe outputs. Therefore, an effective macromodel needs to further compress the number of ports when p_i and p_o are both large. In Section 4 we identify a much smaller number of principal input and output ports by studying the correlation.

3. PROBLEM FORMULATION

3.1 Levelized Via Tracks

Same as [Goplen and Sapatnekar 2005], this paper assumes that the via allocation is after the placement and global routing but before the detailed routing of the signal nets. The active device layers and power/ground planes are first discretized into tiles according to the complexity of the circuit. The inputs of thermal power or switching currents are injected at tiles called input ports. The outputs of temperature gradient and voltage bounce are probed at tiles called output ports. Recall that there are p_o output ports for the electrical-*RLC* model probed at power/ground I/Os connecting the power supply between the package and the bottom device layer.

Moreover, as shown in Fig. 1, because the bottom device layer has the longest path to the heat-sink on the top, the p_o output ports for the thermal- RC model are probed at the bottom device layer in this paper.

The power ground vias are placed at centers of tiles between two layers, and follow an aligned path from the bottom package I/Os to the top heat-sink. We call those aligned paths *vertical tracks* or *tracks*. As vias are aligned, the p_o tracks pass both p_o output ports of the electrical- RLC model and p_o output ports of the thermal- RC model. The density of power/ground vias at each track is the primary design parameter considered in this paper. The density adjusts to satisfy two requirements at output ports. The first is the integrity constraint of the temperature gradient and voltage bounce. The second is the resource constraint with provided signal net congestion.

Devising a method to decide the number of output ports or tracks is a challenging task. On the one hand, if the number of probed output ports is too small, the design would miss those hotspots or violate the signal routing. On the other hand, it is impossible to explore the via density at each track when the number of probed output ports is large. To systematically allocate via, in this paper, we hierarchically decompose the solution space by *levels*:

DEFINITION 1. *Assuming each layer is discretized into N tiles, a level- i allocation is to hierarchically and symmetrically select 4^i tiles located in the center of one subdivision at one layer, and allocate vias aligned in one track between centers of tiles in adjacent device layers.*

The level i is provided by users according to the complexity of the active circuit on one layout. The leveled via-allocation patterns are shown in Fig. 4. A level-0 pattern means allocating vias in the center tile, and a level-1 pattern means allocating vias in each of the 4 partitions. Note that the locations of tiles in one level is uniformly distributed, but the via density varies at different tiles. Our design freedom is the via-density (number of vias in one tile), which is adjusted according to the sensitivity discussed in Section 5.

An i th-level output-port matrix $\mathcal{L}$ in (2) is

$$\mathbf{L} = [l_1, l_2, \dots, l_{p_o}], \quad p_o = 4^i. \quad (5)$$

Each column l_j ($j = 1, \dots, p_o$) has a non-zero entry ‘1’ at j th output port. Obviously, for both the original model and the reduced model, directly probing either power or thermal integrity and further adjusting the via density at all 4^i output ports would be computationally expensive. In Section 4, we resolve this by compressing output ports.

3.2 Co-Optimization of Dynamic Thermal and Power Integrity

To capture the sharp-transition of temperature change as well as the time-accumulated temperature impact, we employ the thermal-integrity integral used in [Yu et al. 2006] as the measure of dynamic thermal integrity at j th ($j = 1, \dots, p_o$) output port:

$$f_j^T = \int_{t_0}^{t_p} \max[y_j(t), T_c] dt = \int_{t_s}^{t_e} [y_j(t) - T_r] dt, \quad (6)$$

with a pulse-width (t_s, t_e) in a sufficient long time-period t_p all in the scale of thermal-constant (ms). $y_j(t)$ is the transient temperature waveform at j th output port, and T_r is the reference temperature.

To further consider the spatial difference of p_o output ports at the bottom device layer, the overall thermal integrity is defined by a normalized summation:

$$f^T = \frac{\sum_{j=1}^{p_o} f_j^T}{t_p^T \cdot p_o}. \quad (7)$$

Such a measure of thermal integrity takes into account both the temporal and spatial variation of temperature. Similarly, the power-integrity integral f_j^V is defined at j th power/ground I/O with reference voltages Vdd and ground, and integrated at the period t_p^V in the scale of electrical-constant (ns). The overall power integrity f^V is defined similarly to f^T for p_o power/ground I/Os with the reference voltage V_r (0 for ground vias and Vdd for power vias).

Accordingly, we have the following problem formulation:

FORMULATION 1. *Given the targeted voltage bounce V_t for p_o output ports at power/ground I/Os, and the targeted temperature gradient T_t for p_o output ports at bottom device layer, the via-allocation problem is to minimize the total via number, such that the temperature gradient f^T is smaller than T_t and the voltage bounce f^V is smaller than V_t .*

Such a via-allocation problem simultaneously driven by power and thermal integrity can be represented by

$$\begin{aligned} \min \quad & \sum_{j=1}^{p_o} n_j \\ \text{s.t.} \quad & f^V \leq V_t, \quad f^T \leq T_t \\ & \text{and } n_{\min} \leq n_j \leq n_{\max} \end{aligned} \quad (8)$$

Note that n_j is the via density at j th track and V_t and T_t are the targeted voltage bounce and temperature gradient. As discussed in Section 5.2, n_j is decided according to the power and thermal sensitivities obtained from the macromodel. As our power/ground vias are allocated after the placement and global routing of signal nets at each active device layer, the densities of those inter-layer signal nets are available to calculate a maximum density $n_{\max}$ for the power/ground vias. In addition, for the sake of the reliability concern for the large current, the via density n_j on the other hand can not be smaller than a minimum density $n_{\min}$. These parameters $(n_{\max}, n_{\min}, V_t, T_t)$ can be estimated and provided by users.

As presented in Section 5, the key to solving (8) is an efficient yet accurate dynamic analysis to evaluate $V(t)$ and $T(t)$ as well as their sensitivities with respect to p_o tracks, calculated at two different time scales, respectively. After obtaining sensitivities for p_o tracks, we decide the via density n_j in a greedy fashion for each level proportional to the sensitivity.

4. PRINCIPAL INPUT AND OUTPUT IDENTIFICATION

As discussed in Section 2, the effectiveness of macromodels diminishes when there are a large number of inputs and outputs. In this paper, we reduce the complexity

Fig. 4. The levelized tracks to add vias (a) level = 0, (b) level = 1 and (c) level = 2. Vias are allocated between each pair of adjacent active device layers from the bottom package to the top heat-sink.

from ports by studying the correlation of input signals and the correlation of output signals. The previous works [Feldmann and Liu 2004; Liu et al. 2005; Li and Shi 2006], however, extract the correlation by directly studying the topology matrix $\mathcal{B}$ and $\mathcal{L}$. As the correlation physically relates to the input and output signals, this paper extracts the correlation by taking samples of ‘snapshots’ at inputs and outputs. With the use of proper-orthogonal-decomposition (POD) or principal-component-analysis (PCA) [Moore 1981; Box et al. 1994; Antoulas et al. 2001; Rathinam and Petzold 2003; Astrid et al. 2007; Ding and Zha 2008], a large number of ports are compressed into a small number of principal ports.

4.1 Correlation Extraction

Since the electrical signals may share the same clock and operate within a similar logic function, their waveforms in time-domain at certain input ports can show a correlation. Similarly, the thermal power may differ significantly between those regions with and without the clock gating, but can be quite similar inside the region with the same mode as inputs have similar duty-cycles over time. We call this phenomenon *input similarity*.

As the input vector

$$\mathbf{I}(t) = [\mathbf{I}_1 \ \mathbf{I}_2 \ \cdots \ \mathbf{I}_{p_i}] \in R^{p_i \times 1}. \quad (9)$$

is usually known during the physical design, they can be represented by taking a set of ‘snapshots’ sampled at $\mathcal{N}$ time-points

$$\begin{bmatrix} \mathbf{I}_1(t_0) & \cdots & \mathbf{I}_1(t_{\mathcal{N}}) \\ \vdots & \ddots & \vdots \\ \mathbf{I}_{p_i}(t_0) & \cdots & \mathbf{I}_{p_i}(t_{\mathcal{N}}) \end{bmatrix} \quad (10)$$

in a sufficient long period $[0, T_p]$. The sampling cycle is in a different time-scale for the thermal-power (ms) and switching-current (ns).

According to the POD analysis [Antoulas et al. 2001; Rathinam and Petzold 2003; Astrid et al. 2007], the similarity can be mathematically described by a correlation matrix (or Grammian). As such, an *input correlation matrix* among p_i input sources

is estimated by a co-variance matrix in the affine form:

$$\mathcal{R} = \frac{1}{\mathcal{N}} \sum_{\alpha=1}^{\mathcal{N}} (\mathbf{I}(t_{\alpha}) - \bar{\mathbf{I}})(\mathbf{I}(t_{\alpha}) - \bar{\mathbf{I}})^T \in R^{p_i \times p_i}. \quad (11)$$

The $\bar{\mathbf{I}}$ is a vector of mean values defined by:

$$\bar{\mathbf{I}} = \frac{1}{\mathcal{N}} \sum_{\alpha=1}^{\mathcal{N}} \mathbf{I}(t_{\alpha}) \quad (12)$$

Usually, the input vector $\mathbf{I}(t)$ is periodic and the waveform in each period can be approximated by the piecewise-linear model. (11) and (12) are then pre-stored in a table after the evaluation.

An *output similarity* is defined for responses at output ports and measured by a *output correlation matrix*. To extract the output correlation matrix that is independent on the inputs, we assume that p_i inputs in the input vector $\mathbf{I}(s)$ are all the unit-impulse source $h(s)$ and define an input-port vector $\mathcal{J}(s)$ by

$$\mathcal{J} = \mathcal{B}\mathbf{I}(s), \in R^{1 \times N}, \quad (13)$$

which has p_i non-zero entries with the unit-value ‘1’. Accordingly, the p_o output responses $y(s)$ are calculated by

$$\begin{aligned} y(s) &= \mathcal{L}^T(\mathcal{G} + s\mathcal{C})^{-1} \mathcal{J} \\ &= [y_1(s) \ y_2(s) \ \cdots \ y_{p_o}(s)] \in R^{p_o \times 1}. \end{aligned} \quad (14)$$

The according output correlation matrix is extracted in the frequency-domain. Similarly, the output signals can be represented by taking a set of ‘snapshots’ sampled at $\mathcal{N}$ frequency points

$$\begin{bmatrix} y_1(s_0) & \cdots & y_1(s_{\mathcal{N}}) \\ \vdots & \ddots & \vdots \\ y_{p_o}(s_0) & \cdots & y_{p_o}(s_{\mathcal{N}}) \end{bmatrix} \quad (15)$$

in a sufficient wide band $[0, s_{max}]$. The s_{max} locates in a low-frequency range for the temperature and in a high-frequency range for the voltage.

A co-variance matrix is defined under the POD analysis [Antoulas et al. 2001; Rathinam and Petzold 2003; Astrid et al. 2007] in frequency-domain as follows

$$R = \sum_{\alpha=1}^{\mathcal{N}} (y(s_{\alpha}) - \bar{y})(y(s_{\alpha}) - \bar{y})^T \in R^{p_o \times p_o} \quad (16)$$

to estimate the correlation matrix among p_o outputs. The $\bar{y}$ is a vector of mean values defined by:

$$\bar{y} = \frac{1}{\mathcal{N}} \sum_{\alpha=1}^{\mathcal{N}} y(s_{\alpha}) \quad (17)$$

Both (16) and (17) can be evaluated and pre-stored in a table as well.

4.2 Ports Compression

Let $\mathcal{V} = [v_1, v_2, \dots, v_K] (\in R^{N \times K})$ as the first K singular-value vectors of the input correlation matrix $\mathcal{R}$, and $\mathcal{W} = [w_1, w_2, \dots, w_K] (\in R^{N \times K})$ as the first K singular-value vectors of the output correlation matrix R . All singular-value vectors are obtained from the singular-value decomposition (SVD) of $(\mathcal{V}, \mathcal{W})$. The order K is determined when the K th singular-value is below a desired threshold. A rank- K matrix P_i can be constructed by $P_i = \mathcal{V}\mathcal{V}^T$, and a rank- K matrix P_o can be constructed by $P_o = \mathcal{W}\mathcal{W}^T$. Note that both $\mathcal{V}$ and $\mathcal{W}$ are orthonormalized, i.e., $\mathcal{I} = \mathcal{V}^T\mathcal{V}$ and $\mathcal{I} = \mathcal{W}^T\mathcal{W}$. As shown in [Antoulas et al. 2001; Rathinam and Petzold 2003; Astrid et al. 2007], the correlation matrix $(\mathcal{R}, R)$ is essentially the solution that minimizes the least-square between the original states $(\mathbf{I}(t), y(s))$ and their rank- K approximations $(P_i \cdot \mathbf{I}(t), P_o \cdot y(s))$.

As a result, both the input signals $\mathbf{I}(t)$ and the output signals $y(s)$ can be approximated by an invariant (or dominant) subspace spanned by the orthonormalized columns of V and W , respectively:

$$\mathbf{I} = \mathcal{V}\mathbf{I}_K, \quad y = \mathcal{W}y_K. \quad (18)$$

Note that in frequency domain, the output response $y(s)$ is obtained by

$$y(s) = \mathcal{L}^T(\mathcal{G} + s_0\mathcal{C})^{-1}\mathcal{B}\mathbf{I}(s). \quad (19)$$

Based on (18) and (19), it leads to the following equivalent system equation

$$y_K(s) = \mathcal{L}_K^T(\mathcal{G} + s\mathcal{C})^{-1}\mathcal{B}_K\mathbf{I}_K(s), \quad (20)$$

where

$$\mathcal{L}_K^T = \mathcal{W}^T\mathcal{L}^T, \quad \mathcal{B}_K = \mathcal{B}\mathcal{V}. \quad (21)$$

Therefore, both the dimensions of $\mathcal{L}$ ($\in R^{N \times p_o}$) and $\mathcal{B}$ ($\in R^{N \times p_i}$) are greatly reduced when $K \ll p_i$ and p_o . We call $\mathbf{I}_K$ and y_K *principal inputs and outputs* identified by *principal input-port and output-port matrices* $\mathcal{B}_K$ and $\mathcal{L}_K$, respectively.

Therefore, we obtain an equivalent system to the original system. Instead of the input/output identified by $\mathbf{I}/y$, the new system is described by the principal input/output by $\mathbf{I}_K/y_K$. Its transfer function is

$$H_K = \mathcal{L}_K^T(\mathcal{G} + s\mathcal{C})^{-1}\mathcal{B}_K. \quad (22)$$

The according Krylov subspace

$$\mathcal{K}(\mathcal{A}, \mathcal{R}_K) = \{\mathcal{A}, \mathcal{A}\mathcal{R}_K, \dots, \mathcal{A}^{q-1}\mathcal{R}_K, \dots\}$$

is composed by moment matrices

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

Similarly, a projection matrix Q_K that contains the Krylov subspace

$$\mathcal{K}(\mathcal{A}, \mathcal{R}_K, q) \subseteq Q_K.$$

can be obtained by the q th-order block-Arnoldi iteration [Odabasioglu et al. 1998]. Because of $K \ll p_i$, the computational cost to construct the projection matrix Q_K is greatly reduced. Moreover, given the same size n for reduced models, the number of matched block moments is increased from $\lfloor n/p_i \rfloor$ to $\lfloor n/K \rfloor$. Therefore,

the identification of principal input further helps build an effective projection matrix Q_K to reduce the size of system matrices $\mathcal{G}$ and $\mathcal{C}$ by

$$\hat{\mathcal{G}} = Q_K^T \mathcal{G} Q_K \quad \hat{\mathcal{C}} = Q_K^T \mathcal{C} Q_K,$$

and further reduce the size of $\mathcal{L}_K$ and $\mathcal{B}_K$

$$\hat{\mathcal{B}}_K = Q_K^T \mathcal{B}_K \quad \hat{\mathcal{L}}_K^T = Q_K^T \mathcal{L}_K^T.$$

Recall that the p_o outputs are not only the place we probe power and thermal integrity, but also the regions in which we plan to add and adjust the via density. The correlated output ports thereby can be clustered together and further parameterized with a common via density in the same cluster. This is illustrated with details in the following Section 5.

5. PARAMETERIZED MACROMODEL OF INTEGRITY AND SENSITIVITY

The changes at outputs, i.e., sensitivities can be similar for those correlated regions. As such, it can guide us to further group those correlated regions into a small number of ‘representative’ regions, called clusters, and further help effectively generate a parameterized macromodel with a cluster-sensitivity. Therefore, we can efficiently adjust the density of power/ground vias based on the cluster-sensitivity and optimize both the power and thermal integrity. In this Section, we employ a spectral analysis to cluster parameters, present a structured and parameterized macromodeling to provide both nominal values and their sensitivities, and discuss a via allocation algorithm while simultaneously considering power and thermal integrity.

5.1 Clustered Parameterization

The paper as far has only discussed the nominal circuit equation and state variables for active device layer, interlayer-dielectrics, power/ground planes, vertical vias for signal nets, and C4 bumps etc. In this part, we show how to consider the parameterized power/ground vias for both providing supply (return-path) and removing heat.

In Section 2, we have shown how to construct a topology matrix to probe power and thermal integrity at output ports. Output ports are defined at the bottom device layers from where the power/ground vias go up to the top heat-sink and go down to the bottom package I/Os. The power/ground vias are aligned to pass layer by layer. When adding and adjusting vias at output ports, the added vias change RC values in the state matrix at different nodes. As a result, the parameterized description of vias needs to reflect the changes in the state matrices $\mathcal{G}$ and $\mathcal{C}$.

For an i th-level allocation with p_o (4^i) vertical tracks, the locations to add vias are fixed and can be described by p_o topological matrices:

$$X_1 = \begin{bmatrix} 1 & -1 & \cdots \\ \vdots & \vdots & \end{bmatrix} \quad X_j = \begin{bmatrix} \vdots & \vdots & \vdots \\ \cdots & -1 & 2 & -1 & \cdots \\ \vdots & \vdots & \vdots \end{bmatrix} \quad X_{p_o} = \begin{bmatrix} \vdots & \vdots \\ \cdots & -1 & 1 \end{bmatrix} \text{ all } \in R^{N \times N} \quad (23)$$

Fig. 5. The topology matrices X_3 and X_7 when allocating vias between nodes 3 and 7.

Fig. 5 shows one simple example of the topological matrix X_3 and X_7 when allocating vias between tile-3 and tile-7 at two layers.

Then, the unit-conductance and capacitance matrices for j th track are:

$$n_j \cdot g_j = n_j \cdot (gX_j), \quad n_j \cdot c_j = n_j \cdot (cX_j),$$

where g and c are conductance and capacitance for one via with unit area a . Note that the via density n_j is related to the via area A_j in the tile by $n_j = A_j/a$, where a is the unit area of via determined by the processing technology. Because conductance and capacitance values are both proportional to the area A , they are implicitly proportional to the via density n_j as well.

Assuming the via density is n_j for j th track and defining a via density vector

$$\mathbf{n} = [n_1, \dots, n_{p_o}], \quad (24)$$

a parameterized state equation can be obtained

$$(\mathcal{G} + s\mathcal{C} + \sum_{j=1}^{p_o} n_j g_j + s \sum_{j=1}^{p_o} n_j c_j) x(\mathbf{n}, s) = \mathcal{B}_K \mathbf{I}_K(s), \quad y_K(\mathbf{n}, s) = \mathcal{L}_K^T x(\mathbf{n}, s). \quad (25)$$

Similar to [Li et al. 2005; Yu et al. 2006; Yu et al. 2006], we expand $x(\mathbf{n}, s)$ in Taylor series with respect to n_j , and introduce a new state variable x_{ap}

$$x_{ap} = [x^{(0)}, x_1^{(1)}, \dots, x_{p_o}^{(1)}]^T. \quad (26)$$

It contains both the nominal response $x^{(0)}$ and its first-order sensitivities $[x_1^{(1)}, \dots, x_{p_o}^{(1)}]$ with respect to p_o parameters $[n_1, \dots, n_{p_o}]$. The overall responses is obtained by

$$x = x^{(0)} + \sum_{j=1}^{p_o} x_j^{(1)}.$$

Substituting (26) in (25), and explicitly matching expanded terms for each n_j up to the first-order, (25) can be reformulated into a parameterized system with augmented dimension

$$(\mathcal{G}_{ap} + s\mathcal{C}_{ap})x_{ap} = \mathcal{B}_{ap} \mathbf{I}_K(s), \quad y_{ap} = \mathcal{L}_{ap}^T x_{ap}, \quad (27)$$

where the state matrices become

$$\mathcal{G}_{ap} = \begin{bmatrix} \mathcal{G} & 0 & \dots & 0 \\ n_1 g_1 & \mathcal{G} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ n_{p_o} g_{p_o} & 0 & \dots & \mathcal{G} \end{bmatrix}, \quad \mathcal{C}_{ap} = \begin{bmatrix} \mathcal{C} & 0 & \dots & 0 \\ n_1 c_1 & \mathcal{C} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ n_{p_o} c_{p_o} & 0 & \dots & \mathcal{C} \end{bmatrix}, \quad (28)$$

the input/output ports become

$$\mathcal{B}_{ap} = [\mathcal{B}_K, 0, \dots, 0]^T, \quad \mathcal{L}_{ap} = [\mathcal{L}_K, \delta n_1 \mathcal{L}_K, \dots, \delta n_{p_o} \mathcal{L}_K]^T, \quad (29)$$

and the output responses become

$$y_{ap} = [y_K^{(0)}, (y_K^{(1)})_1, \dots, (y_K^{(1)})_{p_o}]^T. \quad (30)$$

Note that state matrices $\mathcal{G}_{ap}$ and $\mathcal{C}_{ap}$ constructed in this fashion both show a lower-block triangular structure. The nominal states are in the diagonal blocks and the perturbed states for different allocation-patterns are in the lower off-diagonal blocks. As such, there only factorization cost coming from diagonal blocks. In addition, the output response y_{ap} have two parts: the nominal value, $y_K^{(0)}$, and its first-order sensitivity $(y_K^{(1)})_j$ with respect to each parameter n_j ($j = 1, \dots, p_o$). Obviously, if we treat each n_j at j th track as an independent parameter, the system size will be become too large to analyze. Thanks for the correlation at output ports, we present a spectral clustering below to cluster those correlated tracks and assign same via density for tracks in one cluster and different via density for tracks in different clusters.

ALGORITHM 1: SPECTRAL CLUSTERING
1 Input: Cluster number K , correlation matrix $R \in R^{p_o \times p_o}$, and p_o topology matrices X_j $j = 1, \dots, p_o$
2 <i>Compute</i> normalized Laplacian: $R_L = D^{-1/2} R D^{1/2}$, where $D = \text{diag}(R)$;
3 <i>Compute</i> the first K singular-value vectors $v_1, \dots, v_K$ of R_L ;
4 <i>Let</i> $V = [v_1, \dots, v_K] \in R^{N \times K}$ and $R_C = R_L \cdot V$;
5 <i>Add</i> j th output port into k th cluster if $R_C(j, k)$ is the maximum in j th row;
6 <i>Form</i> P_K by summing all clustered X_j s in k th cluster;
7 Output: new topology matrices P_k ($k = 1, \dots, K$)

Fig. 6. Algorithm 1 for spectral clustering of parameters.

The spectral analysis can efficiently cluster or partition a large-scale graph by analyzing the spectral values (eigen values or singular-values) of the graph [Alpert et al. 1999]. It is related to the POD or PCA analysis [Kannan et al. 2004; Ding and Zha 2008] if we build a *correlation graph* $G(V, E)$ with p_o output ports as nodes, and their correlation as edges weighed by correlation coefficients. For output ports n_i and n_j , an edge is added with the weight $R(i, j)$. Fig. 7 shows the relation between the correlation graph and the correlation matrix.

The overall procedure of the spectral analysis is outlined in Algorithm 1. To improve the numerical stability, the spectral clustering first normalizes R by its diagonal D to construct the Laplacian R_L of R . Similar to the concept of projection based model order reduction, the spectral clustering in this paper finds a proper projection matrix V composed by the first K singular-value vectors $[v_1, \dots, v_K]$ of R_L . V spans the K th-order subspace with the first K dominant singular-value vectors or directions. By projecting the large-dimensioned $R_L \in R^{p_o \times p_o}$ into the subspace spanned by V , R_L is compressed to

$$R_C = R_L \cdot V. \quad (31)$$

Therefore, if each column vector v_j is viewed as a cluster basis, data points in the correlation graph defined by R_L is clustered into K clusters along K directions, with each direction defined by v_j . As such, an output port j ($j = 1, \dots, p_o$) is placed into a cluster k ($k = 1, \dots, K$) if $R_C(j, k)$ is the largest entry in the j th row of R_C . Accordingly, the j th track to place vias is added into k th cluster. Its related topology matrix X_j is summed into a clustered topology matrix P_k ($k = 1, \dots, K$). Assuming m tracks in one cluster with topology matrices $[X_{j1}, X_{j2}, \dots, X_{jm}]$, then the new clustered topology matrix P_k becomes

$$P_k = X_{j1} + X_{j2} + \dots + X_{jm}. \quad (32)$$

In the same cluster k , all tracks have the same via density n_k . Hence only K parameterized via densities need to adjust instead of p_o via densities ($K \ll p_o$). We call those K via densities *principal parameters*.

As a result, the parameterized system can be compressed into

$$(\mathbf{G}_{ap} + s\mathbf{C}_{ap})\mathbf{x}_{ap} = \mathbf{B}_{ap}\mathbf{I}_K(s), \quad \mathbf{y}_{ap} = \mathbf{L}_{ap}^T \mathbf{x}_{ap}, \quad (33)$$

where the state matrices become

$$\mathbf{G}_{ap} = \begin{bmatrix} \mathcal{G} & 0 & \dots & 0 \\ n_1 g_1 & \mathcal{G} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ n_K g_K & 0 & \dots & \mathcal{G} \end{bmatrix}, \quad \mathbf{C}_{ap} = \begin{bmatrix} \mathcal{C} & 0 & \dots & 0 \\ n_1 c_1 & \mathcal{C} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ n_K c_K & 0 & \dots & \mathcal{C} \end{bmatrix}. \quad (34)$$

The $(\mathbf{B}_{ap}, \mathbf{L}_{ap}, \mathbf{x}_{ap}, \mathbf{y}_{ap})$ are in the same forms as $(\mathcal{B}_{ap}, \mathcal{L}_{ap}, x_{ap}, y_{ap})$ but with an augmented dimension compressed from p_o to K . Note that because the dimension of nominal state matrices in the diagonal is still large, the model order reduction is needed to further reduce (33).

The main computational cost of Algorithm 1 is from the step 3, the singular-value-decomposition (SVD) of the correlation matrix. For a dense correlation matrix ($\in R^{p_o \times p_o}$) with the known-rank K , the complexity of SVD is $O(p_o K^2)$. Usually, the correlation matrix can be first sparsified by pruning small entries of each row. This would help reduce part of the computational cost. Moreover, as the correlation is done once and can be repeatedly used, the use of SVD is thereby an attractive approach to compress physical parameters for the design automation. In addition, when compared to the K-means clustering in [Liu et al. 2005], the spectral clustering developed in this paper determines the cluster by the subspace projection. It does not depend on an arbitrary initial condition that may lead to different clustering results caused by the K-means clustering.

Fig. 7. The parameter clustering by clustering nodes on the correlation graph built from the correlation matrix.

5.2 Macromodel based Via Allocation

A q th-order flat projection matrix $\mathbf{Q}_{ap}$ can be obtained from block Arnoldi iteration to reduce (33). To preserve the separated nominal value and its sensitivity during projection, $\mathbf{Q}_{ap}$ is first partitioned into $K + 1$ blocks

$$\mathbf{Q}_{ap} = [Q_0, \underbrace{Q_1, \dots, Q_K}_K], \quad (35)$$

each Q_j ($j = 0, 1, \dots, K$) with size $N \times q \times K$. Then, similar to [Yu et al. 2006; Yu et al. 2006], $\mathbf{Q}_{ap}$ is further structured as follows

$$\mathbf{Q}_{ap} = \begin{bmatrix} Q_0 & & & \\ & Q_1 & & \\ & & \ddots & \\ & & & Q_K \end{bmatrix} \quad (36)$$

Each Q_j ($j = 0, 1, \dots, K$) is further orthonormalized with each other. As such, proven in [Yu et al. 2005; Yu et al. 2006; Yu et al. 2006] the structured reduction by $\mathbf{Q}_{ap}$

$$\tilde{\mathbf{G}}_{ap} = \mathbf{Q}_{ap}^T \mathbf{G}_{ap} \mathbf{Q}_{ap}, \quad \tilde{\mathbf{C}}_{ap} = \mathbf{Q}_{ap}^T \mathbf{C}_{ap} \mathbf{Q}_{ap}, \quad \tilde{\mathbf{B}}_{ap} = \mathbf{Q}_{ap}^T \mathbf{B}_{ap}, \quad \tilde{\mathbf{L}}_{ap} = \mathbf{Q}_{ap}^T \mathbf{L}_{ap} \quad (37)$$

still preserves moments up to the q th-order.

After projected by $\mathbf{Q}_{ap}$, the reduced macromodel in time-domain can be solved by the Backward-Euler (BE) integration with the state-equation below at time-instant t and time-step h

$$\begin{aligned} (\tilde{\mathbf{G}}_{ap} + \frac{1}{h} \tilde{\mathbf{C}}_{ap}) \tilde{\mathbf{x}}_{ap}(t) &= \frac{1}{h} \tilde{\mathbf{C}}_{ap} \tilde{\mathbf{x}}_{ap}(t-h) + \tilde{\mathbf{B}}_{ap} \mathbf{I}_K(t) \\ \tilde{\mathbf{y}}_{ap}(t) &= \tilde{\mathbf{L}}_{ap}^T \tilde{\mathbf{x}}_{ap}(t). \end{aligned} \quad (38)$$

As shown in [Yu et al. 2005; Yu et al. 2006; Yu et al. 2006], the projection by $\mathbf{Q}_{ap}$ preserves the block matrix structure. As a result, reduced $\tilde{\mathbf{G}}_{ap}$ and $\tilde{\mathbf{C}}_{ap}$ both have the same preserved lower-block triangular structure. Therefore, (38) can be efficiently solved by the block-backward-substitution.

Note that the reduced nominal response and its sensitivity are still separated at outputs:

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

The according overall output response $\tilde{\mathbf{y}}$, i.e., the integrity vector of voltage or temperature (V/T) is

$$\tilde{\mathbf{y}}(\mathbf{n}, t) = \tilde{\mathbf{y}}^{(0)}(\mathbf{n}, t) + \tilde{\mathbf{y}}^{(1)}(\mathbf{n}, t). \quad (39)$$

As discussed below, such a structured and parameterized macromodel can be incorporated into a sensitivity-based allocation.

ALGORITHM 2: SENSITIVITY BASED VIA ALLOCATION
1 Input: K principal input ports, K principal output ports, K principal parameters, maximum temperature bound T_{max} , maximum voltage-bounce bound V_{max} , signal-net-congestion bound n_{max} and current-density bound n_{min} 2 Construct structured and parameterized macromodel by (37); 3 Compute nominal voltage(V)/temperature(T) and sensitivity $\mathbf{S}_V/\mathbf{S}_T$ by (38); 4 Check V_{max} and T_{max} constraints for all tiles; 5 Increase the via density $\mathbf{n}$ according to weighted sensitivity $\mathbf{S}$ in the range of (n_{min}, n_{max}) ; 6 Update the structured and parameterized macromodel according to (34); 7 Repeat from Step 3 until Step 4 is satisfied; 8 Output: Via density vector $\mathbf{n}$

Fig. 8. Algorithm 2 for the sensitivity-based via allocation with the use of macromodels.

Since our macromodel provides both nominal values and sensitivities, it can be incorporated in any gradient-based optimization. The overall optimization to solve the problem formulation (8) is outlined in Algorithm 2. Its inputs are two parts. The first is a principal system by (33) with the identified K principal input and output ports from Algorithm 1, and the clustered K principal parameters. The second is the user provided temperature bound T_{max} , voltage bounce bound V_{max} , signal-net congestion bound n_{max} , and current-density bound n_{min} . Then, a structured and parameterized macromodel is built once by (37). Both nominal responses and sensitivities at K principal input-ports for each perturbed allocation-pattern are solved by (38). If the integrity constraints are not satisfied for K principal tracks, the vias density vector $\mathbf{n}$ are increased according to the sensitivity. This process repeats until the integrity constraints are satisfied.

The sensitivity vector $\mathbf{S}$ in Step 3 is a weighted-maximum of normalized voltage-sensitivity vector $\mathbf{S}_V$ and thermal-sensitivity $\mathbf{S}_T$:

$$\mathbf{S} = \max \left[\alpha \cdot \tilde{\mathbf{y}}_T^{(1)} / \|\tilde{\mathbf{y}}_T^{(1)}\|, \quad \beta \cdot \tilde{\mathbf{y}}_V^{(1)} / \|\tilde{\mathbf{y}}_V^{(1)}\| \right] \in R^{p_o \times 1} \quad (40)$$

	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 II. Electrical and thermal constants.

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 III. Dimensions of 3D ICs layers.

where α and β are weights for $\mathbf{S}_T$ and $\mathbf{S}_V$, and the maximum of the normalized sensitivity is selected at each track.

The sensitivity vector is iteratively updated to calculate the new via density vector by

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

till the integrity constraints are satisfied at all principal output ports. Note that γ is an adaptive-controlled step size and geometrically decreases by a factor of 0.99 as the iteration proceeds. Recall that at each step, $\mathbf{n}$ is constrained by the signal-congestion induced bound n_{max} and the current-density induced bound n_{min} .

The overall computational cost of Algorithm 2 is low. This can be illustrated in two folds. Firstly, the use of the principal input/output identification and the principal parameter clustering leads to an efficient evaluation of integrity at principal ports. Secondly, due to the structure reduction, the nominal values and their sensitivities can be efficiently solved by the block-backward substitution of (38), and only the perturbed states are updated during each iteration according to (34).

A formal complexity analysis of Algorithm 2 is illustrated as follows. The main computational cost of Algorithm 2 is from steps 2 and 3, the model order reduction and update of the integrity and sensitivity. A q th-order model order reduction by Krylov-subspace projection needs one LU decomposition, q solves and the orthonormalization. Due to the lower-block triangular structure, the LU factorization is only applied for the nominal matrix in the diagonal. Moreover, there are only K off-diagonal blocks for K principal parameters. As such the total complexities of model order reduction is $O(qN^{1.5} + q(KN)^{1.5} + Nq^2)$ ($x, y \in [0, 9]$). The computational cost of updating integrity and sensitivity is $O(q^3 + Kq^2)$. The model order reduction is done only once and the update is done iteratively until the optimization converges.

6. EXPERIMENT RESULTS

6.1 Settings

The proposed two algorithms have been implemented in C and MATLAB. Experiments are run on a Sun-Fire-V250 workstation with 2G RAM. The via allocation based upon the steady-state analysis [Goplen and Sapatnekar 2005; Cong

Fig. 9. Singular-value distribution of the input correlation matrix. x-axis is the order of the singular values.

and Zhang. 2005] is used as the baseline. The reported number of vias is for power/ground vias. silicon, copper and dielectric are assumed for via, heat-sink, active device layer, inter-layer and PG plane, respectively. Table 2 and Table 3 summarize their electrical and thermal constants and dimensions. During the sensitivity-driven allocation, the maximum step γ is $3e4$ with the damping factor 0.99, the weight α and β are equal (0.5) for (40). We assume that the ambient air has a fixed temperature $40^\circ C$, and the heat-convection resistance is $0.8K/W$. The Vdd is 2.0V. In addition, to consider the impact of temperature to the electrical resistance, after each optimization, we update the resistance in the power delivery according to $R = R_0(1 + 0.0068 \times T)$. The targeted voltage violation V_{max} is 0.2V and the targeted temperature T_{max} is $52^\circ C$. Two modest 3D stackings are assumed for the experiments. One is with 2-device-layer/2-dielectric-layer and the other is with 4-device-layer/4-dielectric-layer. Moreover, there are 1-heat-sink and 2-P/G-plane. We increase the circuit complexity by increasing the complexity of circuits, the number of input sources, the number of discretized tiles, and the number (level) of tracks for adding vias. The correlation of input or output ports is extracted when the number of ports is large (> 50) and is sparsified with a pruning-threshold ($1e-6$).

As for the thermal model at each active device layer, we have the following assumptions as we have no real designed 3D ICs. Our paper assumes the use of MCNC (http://www.cbl.ncsu.edu/pub/Benchmark_dirs/LayoutSynth92) and IBM benchmark (<http://er.cs.ucla.edu/benchmarks/ibm-place/>) circuits with prior known information of placement and global routing for signal nets [Goplen and Sapatnekar 2005; Yu et al. 2006; Yu et al. 2006]. Each circuit is replicated and placed at each device layer, where the signal nets at two adjacent layers are assumed to be aligned to connect with each other through vertical signal vias. The signal-net congestion n_{max} is thereby calculated by estimating the number of signal nets in each tile at one layer. The n_{min} for current density is assumed to be 10. Moreover, instead of using randomly generated thermal power [Goplen and Sapatnekar 2005; Yu et al. 2006; Yu et al. 2006], the thermal-power inputs are obtained from a pre-simulation of the transient power and temperature simulator [Liao et al. 2005]

Fig. 10. Waveform comparisons between the original and the ones using principal input-ports.(a) is the transient waveform of the temperature at one port, and (b) is the transient waveform of the voltage-bounce at one port.

within the floorplan of DEC-alpha micro-architecture. A sequence of six SPEC2000 applications (*art*, *ammp*, *compress*, *equake*, *gcc*, *gzip*) is applied as input and the thermal-power is recorded every 10-million clock cycles. The clock gating is assumed to have a period of 100ms and the power in the standby mode is assumed to be 20% of the running mode. The profile of thermal power is first uniformly divided inside each functional module on one layer, and then counted as inputs for placed circuits at that layer. An industrial package model is employed with the extracted RLCs and locations of input signals. The switching currents are modeled by the triangular waveform with falling time and rising time 5ns. Similarly, the sequences of input currents are generated by simulating the specified input vectors for an evaluation period of millions of clock cycles.

6.2 Principal Ports Identification

The first example is a 3D stacking with 3-level (64) tracks. There are 6K tiles with 100 thermal-power inputs and 200 switching-current inputs. We first show the result of input compression. Fig. 9 shows the difference in magnitude for the singular-values of the correlation matrix for input ports. Clearly, the singular values of the correlation matrix decrease quickly for high orders. The number of principal input-ports can be determined according to the singular values. In this paper, the threshold value is set as 0.1 of the maximum singular value, and its according order is set as the number of principal input-ports. In Fig. 9, the K=8 is selected as such. Similarly, the same procedure is applied to identity principal inputs for the thermal-power. Fig. 10 compares the time-domain waveforms between our macromodels with port-reduction and the exact MNA solution at selected ports for the previous example in one period of thermal model and electrical model, respectively. The input-port matrix of the electrical model is approximated with K=16 and K=8, respectively. Then, a 6th-order projection is applied to generate the electrical macromodel. Similarly, the input-port matrix of the thermal model

Fig. 11. Output port compression. (a) is the clustering results, and (b) is the transient waveforms of the first cluster.

is approximated with $K=10$ and $K=5$, and an 3rd-order projection is applied to generate the thermal macromodel. Clearly, both macromodels are visually identical to the original ones.

Next, we show the result of output compression and clustering. For the electrical-*RLC* network, Fig. 11 (b) shows the clustering result by the spectral clustering, and Fig. 11 (a) shows the waveforms of 5 clustered ports at cluster-3. The singular-value-decomposition (SVD) discovers that the singular values degrade significantly after 4th order. Therefore, the first four singular-value vectors are used to compose the subspace for clustering. The Spectral analysis leads to 4 clusters with 5, 15, 24, 20 clustered ports, respectively. For the 5 clustered output ports at cluster-3, the waveforms are clearly similar with each other with a maximum peak difference by 0.01V. Based on the result of clustered outputs, i.e., tracks, the vias are allocated with uniform density in one cluster and non-uniform density at different clusters.

6.3 Macromodel based Optimization

For the same example above, Fig. 12 presents the successive decrease of the transient temperature and voltage violation during the simultaneous optimization at one of principal output ports. The via number of this example is minimized in 6 iterations with the targeted voltage bounce 0.2V and the targeted temperature $52^{\circ}C$. Clearly, our sensitivity based allocation can effectively and efficiently minimize the via numbers under the dynamic power and thermal constraints.

Fig. 13 further shows the steady-state temperature map across the bottom device layer. In this example, we assume that all thermal-power sources are located along one side of the device layer. The initial chip temperature at the bottom layer is $150^{\circ}C$, and its temperature profile at steady-state is shown in Fig. 13 (a). In contrast, the via-allocation results in a cooler temperature that closely approaches the targeted temperature as shown in Fig. 13 (b). Clearly, even at steady-state the temperature is still spatially variant. An accurate measure of integrity is therefore needed to consider a time and space averaged integrity at selected probing ports.

Fig. 12. Iterative optimizations showing the reduction of (a) temperature and (b) voltage violation by via-allocation.

Fig. 13. Steady-state temperature map of bottom device layer (a) before allocating via, and (b) after allocating via in a different temperature scales.

6.4 Comparison between Steady-state and Transient Thermal Analysis

In the following, we present the scalability study of via allocations. We increase the circuit complexity by the number of cells, input ports and output ports. Input and output compression are both applied when the number of ports is large (> 50). Table 4 summarizes the complexities of the circuits, the reduced sizes and the principal ports.

Fig. 14. Visualized via-distribution. (a) is the result under only the power-integrity constraint, (b) is the result under only the thermal-integrity constraint, and (c) is the result under both of power and thermal integrity constraints simultaneously.

We compare the runtime and the number of vias in Table 5. The runtime includes the integrity and sensitivity analysis time and optimization time. As all optimizations usually converge in about 10 iterations, the optimization time is small and hence the analysis time dominates runtime. The baseline is the via allocation with the sequential optimization of thermal/power integrity using the static integrity [Goplen and Sapatnekar 2005]. In Table 5, column 2-3 show the runtime and allocated via number for the baseline, and column 4-8 show the results for the optimizations using the dynamic integrity. In detail, column 4 shows the runtime of transient analysis using macromodels without the port-compression, and column 5 shows the number of allocated vias under the sequential optimization. Column 6 shows the runtime of transient analysis using macromodels with the port-compression, and column 7-8 shows the number of allocated vias under the sequential and simultaneous optimizations, respectively.

We first compare the baseline with the design utilizing the dynamic analysis. Both methods allocate vias in a sequential fashion. I.e., we first allocate dummy thermal vias to satisfy thermal integrity with a targeted temperature gradient 52°C , and then allocate power/ground vias to satisfy power integrity with a targeted voltage bounce 0.2V . As shown by Table 5, 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 12% for the first example and 17% for the second one on average. This is because our transient thermal analysis can accurately generate the dynamic thermal integrity, but the steady-state analysis has to assume a constant thermal integrity for all tiles.

Furthermore, the use of macromodels reduces the computational cost to solve power and thermal integrity and their sensitivities. Compared to the macromodel without the port-compression, the macromodeling with the port-compression reduces the overall runtime up to 16X with similar allocation results. Compared to the steady-state analysis with the full-matrix analysis, our macromodel with the

ckt	total tile#	reduced size (T,V)	input src# (T,V)	K-input (T,V)	output track#	K-output (T,V)
ckt1(2-layer)	1.9K	(30,80)	(10,20)	(10,20)	4 ²	(4 ² , 4 ²)
ckt2(2-layer)	6K	(15,48)	(100,200)	(5,8)	4 ³	(6,4)
ckt3(2-layer)	12K	(80,160)	(300,600)	(10,16)	4 ⁴	(8,5)
ckt4(2-layer)	27K	(96,180)	(1K,2K)	(12,18)	4 ⁴	(10,8)
ckt5(2-layer)	52K	(96,220)	(1K,3K)	(12,20)	4 ⁵	(12,14)
ckt1(4-layer)	4K	(50,120)	(20,20)	(20,20)	4 ²	(4 ² , 4 ²)
ckt2(4-layer)	12K	(40,64)	(200,200)	(8,8)	4 ³	(6,4)
ckt3(4-layer)	24K	(180,640)	(600,600)	(15,16)	4 ⁴	(8,5)
ckt4(4-layer)	54K	(240,720)	(2K,2K)	(20,18)	4 ⁴	(10,9)
ckt5(4-layer)	100K	(300,1K)	(2K,3K)	(20,20)	4 ⁵	(12,14)

Table IV. The complexity of the original circuits and the reduced circuits including: the size, number of input-ports, and number of output-ports.

ckt	Steady-state(direct)		Transient(MACRO-1)		Transient(MACRO-2)		
	runtime (s)	total via # by seq-opt	runtime (s)	total via # by seq-opt	runtime (s)	total via # by seq-opt	total via # by sim-opt
ckt1(2-layer)	5.4	178800	0.63	153800 (-13%)	0.63	153800 (-13%)	112800 (-36%)
ckt2(2-layer)	29.7	184900	0.81	159600 (-13%)	0.56	159600 (-13%)	118200 (-36%)
ckt3(2-layer)	182.2	218100	18.6	183800 (-16%)	4.2	184200 (-15%)	136200 (-38%)
ckt4(2-layer)	1269.2	234800	165.7	199000 (-15%)	10.3	199600 (-15%)	145600 (-38%)
ckt5(2-layer)	NA	NA	NA	NA	41.2	208600 (NA)	154200 (NA)
ckt1(4-layer)	12.8	219200	1.9	181900 (-17%)	1.9	181900 (-17%)	120500 (-45%)
ckt2(4-layer)	73.8	249000	4.1	206400 (-18%)	1.7	206600 (-18%)	137100 (-45%)
ckt3(4-layer)	609.2	279300	93.7	225800 (-19%)	19.2	226200 (-19%)	150800 (-46%)
ckt4(4-layer)	3308.7	302900	512.3	244600 (-19%)	30.6	245300 (-19%)	164000 (-46%)
ckt5(4-layer)	NA	NA	NA	NA	241.9	269200 (NA)	180100 (NA)

Table V. 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. Two macromodels are used during the transient analysis. Macromodel-1 does not use the port-compression, and macromodel-2 uses the port-compression.

port-compression has a 127X smaller runtime. And the steady-state analysis can not complete the largest example in a reasonable runtime. The maximum transient-waveform difference introduced by the macromodel is about 7% when compared to the exact transient waveform.

6.5 Comparison between Sequential and Simultaneous Optimizations

We further compare the sequential thermal/power optimization with the simultaneous thermal/power optimization. Here both methods allocate vias with the use of dynamic integrity. For the 2-layer example and the 4-layer example in Table 5, our simultaneous optimization reduces the via-cost up to 34% and 44.5% when compared to the sequential optimization with static integrity, and up to 22% and 27.5% when compared to the sequential optimization with dynamic integrity. This demonstrates that the reusing of power/ground vias can reduce the via cost compared to allocate the dummy thermal vias separately from the power/ground vias.

Fig. 14 further visualizes the results of circuit-4 (27K, 2-layer) in Table 5. As we divide a plane into a 2-dimensional output port array (16×16), the x and y-axes are the number of ports along the horizontal and vertical dimensions. (a) shows the distribution of the vias with 8 clusters when only using power integrity constraint,

(b) shows the distribution of the vias with 10 clusters when only using thermal integrity constraint, and (c) shows the distribution of the vias when simultaneously using thermal and power integrity constraints. Clearly, thermal power inputs and switching current inputs lead to a difference distribution of vias. Due to the use of clustering, the vias are allocated with uniform density in tracks at one cluster and non-uniform density for different clusters.

7. CONCLUSIONS

The existing work on via-allocation does not consider thermal and power integrity simultaneously in 3D ICs. The wise assignment of power/ground vias is effective to reduce the size of the loop current and hence reduce the voltage bounce induced by the inductive coupling. By further extending power/ground vias to the top heat-sink, they can be reused to remove the heat. An allocation of power/ground vias is presented in this paper with simultaneous consideration of the dynamic power and thermal integrity. Such a problem formulation can significantly avoid an over-design when allocating vias in 3D ICs.

To efficiently solve the problem with dynamic-integrity constraints, two techniques, the compression of ports (parameters) and the structured and parameterized macromodeling are developed to effectively and efficiently generate both of the integrity and its sensitivity. Compared to the baseline of the existing approach using the steady-state thermal analysis and the sequential optimization, experiments show that our simultaneous optimization of power and thermal integrity reduces non-signal vias up to 45.5%.

ACKNOWLEDGMENTS

This work is partially supported by NSF CCR-0401682, and a UC-MICRO grant sponsored by Intel and Mindspeed. In addition, the authors thank the reviewers for insightful comments to make the paper better.

REFERENCES

- ALPERT, C., KAHNG, A., AND YAO, S. 1999. Spectral partitioning with multiple eigenvectors. *Discrete Applied Mathematics*.
- ANTOULAS, A., SORESENSEN, D., AND GUGERCIN, S. 2001. A survey of model reduction methods for large scale systems. *Contemporary Mathematics*, 193–219.
- ASTRID, P., WEILAND, S., AND WILLCOX, K. 2007. Missing point estimation in models described by proper orthogonal decomposition. *IEEE Trans. Autom. Control*.
- BANERJEE, K., SOURI, S. J., KAPUR, P., AND SARASWAT, K. C. 2001. 3D ICs: A novel chip design for improving deep submicron interconnect performance and systems-on-chip integration. *Proc. IEEE*, 602–633.
- BOX, G., JENKINS, G., AND REINSEL, G. 1994. Time series analysis: Forecasting and control. In *Prentice Hall, New York*.
- CHEN, J. AND HE, L. ISPD-2006. Noise-driven in-package decoupling capacitance insertion. In *Proc. Int. Symp. on Physical Design*.
- CHIANG, T.-Y., BANERJEE, K., AND SARASWAT, K. C. ICCAD-2001. Compact modeling and spice-based simulation for electrothermal analysis of multilevel ulsi interconnects. In *Proc. Int. Conf. on Computer Aided Design*.
- CONG, J., WEI, J., AND ZHANG, Y. ICCAD-2004. A thermal-driven floorplanning algorithm for 3d ics. In *Proc. Int. Conf. on Computer Aided Design*.

- CONG, J. AND ZHANG., Y. ICCAD-2005. Thermal via planning for 3D ICs. In *Proc. Int. Conf. on Computer Aided Design*.
- DAS, S. 2004. *Design Automation and Analysis of Three Dimensional Integrated Circuits (Ph. D Thesis)*. Massachusetts Institute of Technology.
- DAVIS, W. AND ET AL. 2005. Demystifying 3D ICs: the pros and cons of going vertical. *IEEE Design and Test of Computers*, 498–510.
- DING, C. AND ZHA, H. 2008. *Spectral Clustering, Ordering and Ranking -Statistical Learning with Matrix Factorizations*. Springer.
- E.J.GRIMME. 1997. *Krylov projection methods for model reduction (Ph. D Thesis)*. Univ. of Illinois at Urbana-Champaign.
- FELDMANN, P. AND LIU, F. ICCAD-2004. Sparse and efficient reduced order modeling of linear sub-circuits with large number of terminals. In *Proc. Int. Conf. on Computer Aided Design*.
- GOPLIN, B. AND SAPATNEKAR, S. ICCAD-2003. Efficient thermal placement of standard cells in 3d ics using a force directed approach. In *Proc. Int. Conf. on Computer Aided Design*.
- GOPLIN, B. AND SAPATNEKAR, S. ISPD-2005. Thermal via placement in 3D ICs. In *Proc. Int. Symp. on Physical Design*.
- HONG, Y. AND ET. AL. ISQED-2005. Analysis for complex power distribution networks considering densely populated vias. In *Proc. Int. Symp. on Quality Electronics and Design*.
- KANNAN, R., VEMPALA, S., AND VETTA, A. 2004. On clusterings: good, bad and spectral. *Journal of the ACM*.
- KNICKERBOCKER, J. U. AND ET. AL. 2005. Development of next-generation system-on-package technology based on silicon carriers with fine-pitch chip interconnection. *IBM Research Journal (Power and Packaging)*, 725–754.
- LI, H. AND ET. AL. DAC-2005. Partitioning-based approach to fast on-chip decap budgeting and minimization. In *Proc. Design Automation Conf.*
- LI, P., PILEGGI, L., ASHEGHI, M., AND CHANDRA, R. ICCAD-2004. Efficient full-chip thermal modeling and analysis. In *Proc. Int. Conf. on Computer Aided Design*.
- LI, P. AND SHI, W. DAC-2006. Model order reduction of linear networks with massive ports via frequency-dependent port packing. In *Proc. Design Automation Conf.*
- LI, X., LI, P., AND PILEGGI, L. Li-2005. 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*.
- LI, Z., HONG, X., ZHOU, Q., YANG, H., V.PITCHUMANI, AND CHENG, C. ISPD-2006. Integrating dynamic thermal via planning with 3D floorplanning algorithm. In *Proc. Int. Symp. on Physical Design*.
- LIAO, W., HE, L., AND LEPAK, K. 2005. Temperature and supply voltage aware performance and power modeling at microarchitecture level. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 1042–1053.
- LIM, S. 2005. Physical design for 3D system-on-package: Challenges and opportunities. *IEEE Design and Test of Computers*, 532–539.
- LIU, P., TAN, S. X.-D., KONG, J., MCGAUGHY, B., AND HE, L. ICCAD-2005. Efficient method for terminal reduction of interconnect circuits considering delay variations. In *Proc. Int. Conf. on Computer Aided Design*.
- MOORE, B. 1981. Principal component analysis in linear systems: controllability, observability, and model reduction. *IEEE Trans. Autom. Control*, 17–32.
- ODABASIOGLU, A., CELIK, M., AND PILEGGI, L. 1998. PRIMA: Passive reduced-order interconnect macro-modeling algorithm. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 645–654.
- RATHINAM, M. AND PETZOLD, L. 2003. A new look at proper orthogonal decomposition. *SIAM Journal on Numerical Analysis*, 1893–1925.
- RUEHLI, A. E. 1974. Equivalent circuits models for three dimensional multiconductor systems. *IEEE Trans. on Microwave Theory and Techniques*, 216–220.

- RYU, C. AND ET. AL. EPEP-2005. 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)*.
- SAPATNEKAR, S. ICICDT-2006. Physical design automation challenges for 3d ics. In *Int. Conf. on IC Design and Tech.*
- SU, H., K.GALA, AND SAPATNEKAR, S. 2003. Analysis and optimization of structured power/ground networks. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 1533–1544.
- TENG, C. C., CHENG, Y. K., ROSENBAUM, E., AND KANG, S. M. 1997. iTEM: A temperature-dependent electromigration reliability diagnosis tool. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 882–893.
- TIWARI, V., SINGH, D., RAJGOPAL, S., MEHTA, G., PATEL, R., AND BAEZ, F. DAC-1998. Reducing power in high-performance microprocessors. In *Proc. Design Automation Conf.*
- WANG, T. AND CHEN, C. 2003. 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*, 691–700.
- YU, H., HE, L., AND TAN, S. BMAS-2005. Block structure preserving model reduction. In *IEEE International Workshop on Behavioral Modeling and Simulation (BMAS)*.
- YU, H., HO, J., AND HE, L. ICCAD-2006. Simultaneous power and thermal integrity driven via stapling in 3D ICs. In *Proc. Int. Conf. on Computer Aided Design*.
- YU, H., SHI, Y., HE, L., AND KARNIK, T. ISLPED-2006. Thermal via allocation for 3D ICs considering temporally and spatially variant thermal power. In *Int. Symp. on Low Power Electronics and Design (ISLPED)*.
- ZHAN, Y. AND SAPATNEKAR, S. S. 2007. High efficiency green function-based thermal simulation algorithms. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*.
- ZHAO, J., ZHANG, J., AND FANG, J. EPEP-1998. Effects of power/ground via distribution on the power/ground performance of C4/BGA packages. In *Proc. IEEE Electrical Performance of Electronic Packaging (EPEP)*.
- ZHAO, M. AND ET.AL. DAC-2007. On-chip decoupling capacitance and p/g wire co-optimization for dynamic noise. In *Proc. Design Automation Conf.*
- ZHAO, M., PANDA, R., AND ET.AL. 2002. Hierarchical analysis of power distribution networks. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 159–168.
- ZHAO, S., ROY, K., AND KOH, C. 2002. Decoupling capacitance allocation and its application to power supply noise aware floorplanning. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 81–92.
- ZHENG, H., KRAUTER, B., AND PILEGGI, L. CICC-2003. On-package decoupling optimization with package macromodels. In *Proc. IEEE Custom Integrated Circuits Conference*.