

Dynamic Power and Thermal Integrity in 3D Integration

Hao Yu and Lei He

Abstract— This paper presents the state-of-art advance in 3D integrations. It illustrates the need of a high-performance 3D design driven by the dynamic power and thermal integrity. The through-silicon-via (TSV) is used to simultaneously deliver the power supply and remove the heat. More importantly, to scope with the large-scale design complexity, the modern macromodeling technique is applied to handle not only the large number of dynamic inputs/working-loads but also the large of size of RLC/RC networks distributing the power/heat. Experiment showed promising results to apply the 3D design presented by this paper.

I. INTRODUCTION

The high-performance VLSI integration by the technology scaling has confronted with the dramatically increased design cost from the noise, power and process variation. The emergence of the multi-core system integration is becoming an alternative design paradigm to improve the performance by increasing the throughput rate. However, in today's two dimensional (2D) Systems-on-chip (SoC) integration, the memory is surrounded by logic circuits and its performance in terms of memory bandwidth is limited by the long interconnect length. Thanks to the recent advance in the three dimensional (3D) integration [1]–[7], a 3D integration can reduce the physical distance between the memory and logic circuits and hence has shown a promising potential to integrate hundreds of cores with a better scaled performance than the 2D integration.

Since there are large numbers of devices densely packed in a number of device layers, it brings a significant burden for the heat removal and power (supply voltage) delivering in 3D ICs. This paper discusses an allocation of through-silicon-via (TSV) to simultaneously consider the dynamic power and thermal integrity in 3D ICs. In Section II, we first illustrate the need of dynamic power and thermal integrity in 3D design, and present a TSV allocation problem and the challenge to solve this problem. In Section III, we discuss how to apply the modern macromodeling technique to reduce the design complexity, and present an efficient solution for our TSV allocation. We show the result in Section IV and conclude the paper in Section V.

II. HIGH-PERFORMANCE 3D DESIGN

A. Dynamic Power and Thermal Integrity

Fig. 1 illustrates a typical 3D stacking of multiple device layers within one package. The supply voltage is deliv-

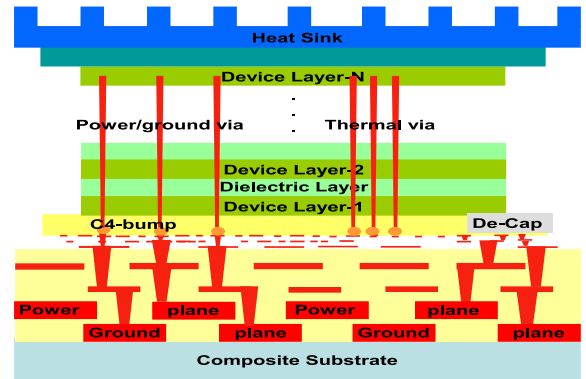


Fig. 1. A typical 3D stacking with non-signal through-vias.

ered 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 to deliver the supply voltage as *power/ground vias*. The 3D integration, by definition, has integrated more than one layer of the active device. They draw much larger current from package power/ground planes than 2D ICs. This can obviously result in the IR drop for the horizontal on-chip power/ground grid. The surge of the injecting current further leads 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. They form a number of different sized loop-inductances that have significant couplings with each other. We call the voltage bounce at I/Os as *power integrity* in this paper.

On the other hand, due to the increased power density and the slow heat-convection at inter-layer dielectrics, the heat dissipation is another concern in 3D ICs [1]. The excessively high temperature can significantly degrade the reliability and performance of interconnects and devices [1], [2], [4], [6]–[9]. We call the temperature gradient at active device layers as *thermal integrity*. 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. One observation is that there are through vias delivering supply voltages or signals from the bottom package through each active device layer. Since the 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 [9] to reduce effective thermal resistances. Its physical arrangement is further studied in [2], [4], [6], [7].

This work was supported by UC-Micro fund from Intel and Mindspeed. Hao Yu is with Berkeley Design Automation, Santa Clara, CA 95054, USA hao.yu@berkeley-da.com. Lei He is with the Department of Electrical Engineering, UCLA, Los Angeles, CA 90095, USA lhe@ee.ucla.edu.

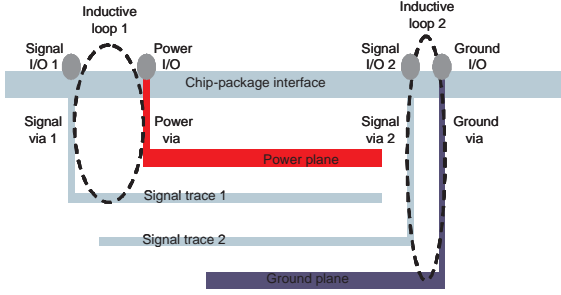


Fig. 2. The power delivery by vertical power/ground vias and its impact on inductive current loops.

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 inputs are the time-varying thermal power (See Fig. 3) [10], [11] 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 [6]. As a result, 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 is thereby needed to accurately guide the physical level resource allocation. 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 [6]–[9], 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 the static thermal-integrity based design [4], [12]. Note that the dynamic power-integrity has already been employed in many on-chip or off-chip power integrity verifications and designs [13]–[15]. A similar *dynamic power integrity* 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.

B. TSV Allocation Problem

We notice that the previous thermal via allocations [4], [6], [12] assume adding dummy vias to conduct heat. They ignore the fact that power/ground vias can help heat-removal

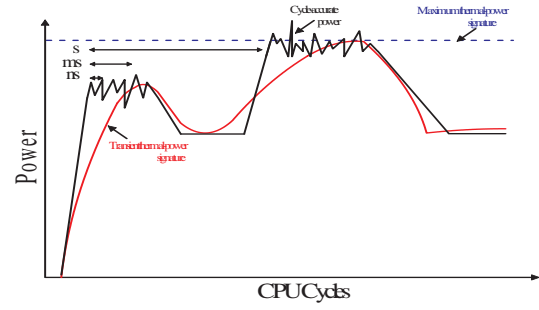


Fig. 3. The definitions of the cycle-accurate power, transient thermal-power, and maximum thermal-power at the different scale of time-constant.

as well. Therefore, the reusing of the power/ground via as thermal via can save the routing resource for signal nets. More importantly, the allocation of power/ground vias can 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.

Same as [12], this paper assumes that the via allocation is after the placement and global routing but before the detailed routing of the signal nets. 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.

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

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

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. f^V and f^T are the metrics of power integrity and thermal integrity, defined by the spatially-averaged time-integral of the transient $V(t)$ [15] and $T(t)$ [6], respectively.

As discussed later in Section III, 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.

III. INTEGRITY OPTIMIZATION WITH MACROMODEL

We represent eh 3D ICs by two distributed models, a thermal- RC model for the heat-removal and an electrical- RLC model for the power-delivery. They (without power/ground vias) can be described in the state-space by

$$\mathcal{G}x(t) + \mathcal{C} \frac{dx(t)}{dt} = \mathcal{B}I(t), \quad y(t) = \mathcal{L}^T x(t) \quad (2)$$

or in frequency (s) domain

$$(\mathcal{G} + s\mathcal{C})x(s) = \mathcal{B}I(s), \quad y(s) = \mathcal{L}^T x(s). \quad (3)$$

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.

During the design optimization of our TSV allocation problem, the main difficulties to directly apply the above state-space equation come from three-fold. Firstly, there are so many inputs to try and so many outputs to probe. Secondly, the dimension of the distributed thermal- RC and electrical- RLC models are too large to analyze. Thirdly, for the sake of design optimization, we are more interested in the sensitivity than the nominal response. In the following, we will show how to compress the I/Os and further generate an effective structure and parameterized macromodels for the design automaton.

A. Compression of I/Os

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. 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. Based on the correlation, we can reduce the redundancy in I/Os by identifying those principal ports.

We call this phenomenon *input similarity*. As the input vector

$$\mathbf{I}(t) = [\mathbf{I}_1 \quad \mathbf{I}_2 \quad \cdots \quad \mathbf{I}_{p_i}] \in R^{p_i \times 1}. \quad (4)$$

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

time-points

$$\begin{bmatrix} \mathbf{I}_1(t_0) & \cdots & \mathbf{I}_1(t_{\mathcal{N}}) \\ \vdots & \ddots & \vdots \\ \mathbf{I}_{p_i}(t_0) & \cdots & \mathbf{I}_{p_i}(t_{\mathcal{N}}) \end{bmatrix} \quad (5)$$

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 [16], the similarity can be mathematically described by a correlation matrix (or Grammian), estimated by a co-variance matrix:

$$\mathcal{R} = \frac{1}{\mathcal{N}} \sum_{\alpha=1}^{\mathcal{N}} (\mathbf{I}(t_{\alpha}) - \bar{\mathbf{I}})(\mathbf{I}(t_{\alpha}) - \bar{\mathbf{I}})^T \in R^{p_i \times p_i}. \quad (6)$$

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

$$\bar{\mathbf{I}} = \frac{1}{\mathcal{N}} \sum_{\alpha=1}^{\mathcal{N}} \mathbf{I}(t_{\alpha}) \quad (7)$$

Usually, the input vector $\mathbf{I}(t)$ is periodic and the waveform in each period can be approximated by the piecewise-linear model.

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

$$\mathcal{J} = \mathcal{B}\mathbf{I}(s), \quad \in R^{1 \times \mathcal{N}}, \quad (8)$$

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

$$\begin{aligned} y(s) &= \mathcal{L}^T (\mathcal{G} + s\mathcal{C})^{-1} \mathcal{J} \\ &= [y_1(s) \quad y_2(s) \quad \cdots \quad y_{p_o}(s)] \in R^{p_o \times 1}. \end{aligned} \quad (9)$$

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

$$\begin{bmatrix} y_1(s_0) & \cdots & y_1(s_{\mathcal{N}}) \\ \vdots & \ddots & \vdots \\ y_{p_o}(s_0) & \cdots & y_{p_o}(s_{\mathcal{N}}) \end{bmatrix} \quad (10)$$

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 in frequency-domain as follows

$$R = \sum_{\alpha=1}^{\mathcal{N}} (y(s_{\alpha}) - \bar{y})(y(s_{\alpha}) - \bar{y})^T \in R^{p_o \times p_o} \quad (11)$$

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

$$\bar{y} = \frac{1}{\mathcal{N}} \sum_{\alpha=1}^{\mathcal{N}} y(s_{\alpha}) \quad (12)$$

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 $(\mathcal{V}, \mathcal{W})$. 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 $P_o = \mathcal{W}\mathcal{W}^T$. As shown in [16], 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:

$$\mathbf{I} = \mathcal{V}\mathbf{I}_K, \quad y = \mathcal{W}y_K. \quad (13)$$

Based on (13), it leads to the following equivalent system equation

$$(\mathcal{G} + s\mathcal{C})x_K(s) = \mathcal{B}_K\mathbf{I}_K(s), \quad y_K(s) = \mathcal{L}_K^T x_K(s) \quad (14)$$

where

$$\mathcal{L}_K^T = \mathcal{W}^T \mathcal{L}^T, \quad \mathcal{B}_K = \mathcal{B}\mathcal{V}. \quad (15)$$

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.

B. Dynamic Sensitivity by Structured and Parameterized Macromodel

Recall that the design parameter in our problem formulation is the the via density at one track. Blindly allocate the via by searching all kinds of combinations would be computationally expensive if not impossible. We decide the via density based on the changes at outputs, i.e., sensitivities, caused by the change of via density.

Let's first parameterize the nominal system 3. The added via is described by two parameters: n_j the via density and X_j the topological matrix that connects the via into the nominal system. As such, a parameterized state-space description can be obtained by

$$\begin{aligned} (\mathcal{G} + s\mathcal{C} + \sum_{j=1}^{p_o} n_j g_j + s \sum_{j=1}^{p_o} n_j c_j)x(\mathbf{n}, s) &= \mathcal{B}_K\mathbf{I}_K(s), \\ y_K(\mathbf{n}, s) &= \mathcal{L}_K^T x(\mathbf{n}, s). \end{aligned} \quad (16)$$

Similar to [6], [7], [15], we expand $x(\mathbf{n}, s)$ in Taylor series with respect to n_j , and introduce a new state variable x_{ap}

$$x_{ap} = [x^{(0)}, x_1^{(1)}, \dots, x_{p_o}^{(1)}]^T. \quad (17)$$

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

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

	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 I
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 II
DIMENSIONS OF 3D ICs LAYERS.

Substituting (17) in (16), (16) can be reformulated into a parameterized system with augmented dimension by

$$(\mathcal{G}_{ap} + s\mathcal{C}_{ap})x_{ap} = \mathcal{B}_{ap}\mathbf{I}_K(s), \quad y_{ap} = \mathcal{L}_{ap}^T x_{ap}, \quad (18)$$

where \mathcal{G}_{ap} and \mathcal{C}_{ap} show a lower-triangular-block structure and hence x_{ap} can be solved from block-backward-substitution.

To further compress the dimension of the state-matrices \mathcal{G}_{ap} and \mathcal{C}_{ap} , we first construct a lower-dimensional subspace Q_{ap} from the moment expansion of (18), and then transform Q into the block-diagonal form Q_{ap} . After the block-orthonormalization of Q_{ap} , we apply a two-side projection to (18) by Q_{ap} and obtain a dimensioned-reduced system with preserved lower-triangular-block structure [6], [7], [14], [15]. The accuracy of macromodel is preserved to match the dominant moments of the original model. More importantly, due to the structure-preserving, both of the nominal response and the sensitivity with regard to the via-density change, can be calculated simultaneously. As such, we can easily embed such a structured and parameterized macromodel into the optimization flow of our TSV allocation problem.

IV. RESULTS

Experiments are implemented in C and MATLAB and run on a Sun-Fire-V250 workstation with 2G RAM. 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 [4], [12]. We use the sequential optimization with the static integrity as the baseline, in comparison to the sequential optimization with the dynamic integrity and the simultaneous optimization with the dynamic integrity proposed in this paper. Table I and Table II summarize the used electrical and thermal constants and dimensions. The targeted voltage violation V_t is 0.2V and the targeted temperature T_t is 52°C. One modest 3D stackings is assumed with 2-device-layer/2-dielectric-layer. Moreover, there are 1-heat-sink and 2-P/G-plane.

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)

TABLE III

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.

We compare the runtime and the number of vias in Table III. In Table III, 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.

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 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.

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. Our simultaneous optimization reduces the via-cost up to 34% when compared to the sequential optimization with static integrity, and up to 22% 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.

V. CONCLUSION

This paper explains the need of dynamic power and thermal integrity for the high-performance 3D integration, using an example of the through-silicon-via (TSV) allocation. To cope with design complexity, an effective macromodel is employed to abstract the physical level detail for the system level design. It composes of the I/O compression and structured and parameterized model order reduction, which efficiently calculate power/thermal integrity and their sensitivity with respect to the via density. Compared to

the design without using the dynamic integrity, experiments show that our approach reduces the the number of TSVs up to 45.5% yet with hundreds of times speedup.

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] B. Goplen and S. Sapatnekar, "Efficient thermal placement of standard cells in 3d ics using a force directed approach," in *Proc. Int. Conf. on Computer Aided Design, ICCAD-2003*.
- [3] S. Das, *Design Automation and Analysis of Three Dimensional Integrated Circuits (Ph. D Thesis)*. Massachusetts Institute of Technology, 2004.
- [4] J. Cong, J. Wei, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3d ics," in *Proc. Int. Conf. on Computer Aided Design, ICCAD-2004*.
- [5] 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.
- [6] H. Yu, Y. Shi, L. He, and T. Karnik, "Thermal via allocation for 3D ICs considering temporally and spatially variant thermal power," in *Int. Symp. on Low Power Electronics and Design (ISLPED)*, ISLPED-2006.
- [7] H. Yu, J. Ho, and L. He, "Simultaneous power and thermal integrity driven via stapling in 3D ICs," in *Proc. Int. Conf. on Computer Aided Design, ICCAD-2006*.
- [8] C. C. Teng, Y. K. Cheng, E. Rosenbaum, and S. M. Kang, "iTEM: A temperature-dependent electromigration reliability diagnosis tool," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 882–893, 1997.
- [9] T.-Y. Chiang, K. Banerjee, and K. C. Saraswat, "Compact modeling and spice-based simulation for electrothermal analysis of multilevel ulsi interconnects," in *Proc. Int. Conf. on Computer Aided Design, ICCAD-2001*.
- [10] 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*.
- [11] W. Liao, L. He, and K. Lepak, "Temperature and supply voltage aware performance and power modeling at microarchitecture level," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 1042–1053, 2005.
- [12] B. Goplen and S. Sapatnekar, "Thermal via placement in 3D ICs," in *Proc. Int. Symp. on Physical Design, ISPD-2005*.
- [13] S. Zhao, K. Roy, and C. Koh, "Decoupling capacitance allocation and its application to power supply noise aware floorplanning," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, pp. 81–92, 2002.
- [14] 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*.
- [15] H. Yu, C. Chu, and L. He, "Off-chip decoupling capacitor allocation for chip package co-design," in *Proc. Design Automation Conf., DAC-2007*.
- [16] P. Astrid, S. Weiland, and K. Willcox, "Missing point estimation in models described by proper orthogonal decomposition," *IEEE Trans. Autom. Control*, 2007.