Resource Allocation Methodology for Through Silicon Vias and Sleep Transistors in 3D ICs

Hailang Wang and Emre Salman
Department of Electrical and Computer Engineering, Stony Brook University,
Stony Brook, NY 11794 USA
E-mail: {hailang.wang, emre.salman}@stonybrook.edu

Abstract—A methodology and analytic expressions are proposed to appropriately allocate the available physical area to through silicon vias (TSVs) and sleep transistors in three-dimensional (3D) ICs with power gating. Power supply noise is minimized by the proposed resource allocation methodology while satisfying the required constraints on leakage current and turn-on time. A comprehensive simulation setup of a three plane 3D IC is developed to evaluate the accuracy and efficacy of the proposed methodology. The proposed expressions exhibit an error of 4% as compared to simulation results. The simulation results also demonstrate that the power supply noise is reduced by more than 46% while satisfying both turn-on time and leakage current.

Keywords—3D IC, TSV, sleep transistor, power gating

I. Introduction

Through silicon via (TSV) based three-dimensional (3D) integration is a promising technology that can address some of the critical issues encountered in traditional 2D ICs such as the adverse effects of global interconnects and limited integration capability [1]. In TSV based vertical integration technologies, multiple wafers are thinned, aligned, and vertically bonded [2].

High density TSVs achieve data communication, power/ground, and clock distribution among the dies and are manufactured using via-first, via-middle, or via-last fabrication methods [3]. The number of TSVs in high performance ICs exceeds 50,000 where the diameter of a single TSV is in the micrometer range [4]. Furthermore, a keep-out zone exists surrounding each TSV to ensure reliable transistor operation. Thus, TSVs consume nonnegligible silicon area in addition to causing routing blockages, as in via-last TSVs.

Due to high levels of integration and existence of heterogeneous blocks, 3D ICs are expected to be heavily power gated to maintain reasonable static current. Power gating requires a large number of sleep transistors where the overall length can exceed one meter [5]. Both power/ground TSVs and sleep transistors significantly affect systemwide power integrity, a primary physical design challenge in 3D ICs. Existing research efforts have investigated TSV types, placement, optimization, and power grid architectures [3, 6, 7]. Decoupling capacitors in 3D ICs have also been considered [8,9]. Power gating in 3D ICs, however, has not received much attention. Wang *et al.* have proposed two power gating topologies each tailored to the specific physical characteristics of via-first and via-last TSVs [10]. In [11], Todri *et al.* have investigated the effect of plane-level power gating on power integrity in 3D ICs.

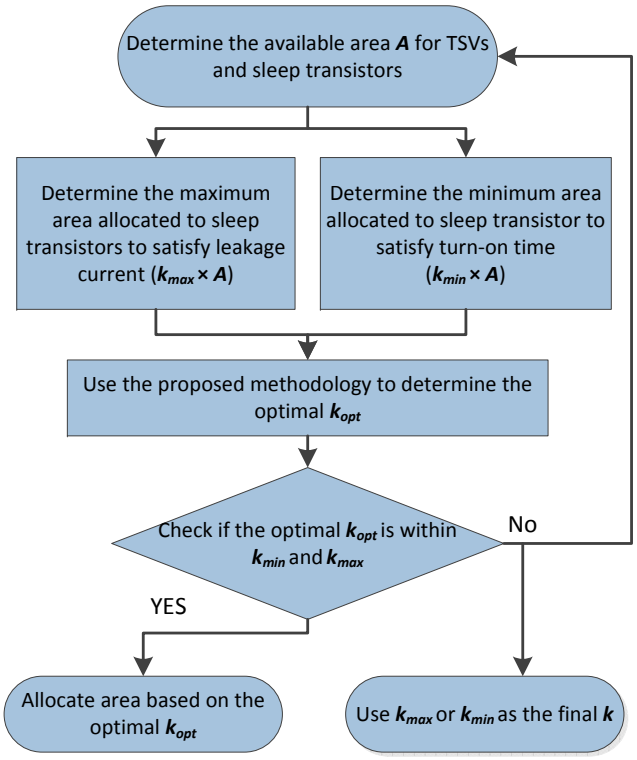

Fig. 1. Summary of the proposed resource allocation methodology.

A methodology and analytic expressions are proposed in this paper to minimize power supply noise by appropriately allocating the available physical area to power/ground TSVs and sleep transistors, while satisfying the constraints on turn-on time and leakage current. A comprehensive simulation setup is developed to evaluate the accuracy and efficacy of the proposed methodology.

The rest of the paper is organized as follows. The proposed methodology and analytic expressions are described in Section II. The simulation framework and results are presented in Section III. Finally, the paper is concluded in Section IV.

This research is supported in part by the National Science Foundation CAREER grant under No. CCF-1253715 and the Office of the Vice President for Research at Stony Brook University.

II. Proposed Methodology

The flow of the proposed methodology is presented in Section II-A. Analytic expressions are developed in Section II-B to determine the optimal area allocation. The dependence of optimal allocation on available area is discussed in Section II-C.

A. Summary of the Proposed Flow

The proposed resource allocation method is summarized in Fig. 1. The first step is to determine the available physical area A for power/ground TSVs and sleep transistors. Assume that the ratio of this area allocated to sleep transistors is k [*i.e.* sleep transistors occupy an area of $k \times A$ whereas TSVs occupy an area of $(k-1) \times A$]. Next, the minimum and maximum sleep transistor sizes are determined to satisfy the constraints on, respectively, turn-on time and leakage current. If less area is allocated to sleep transistors than the minimum required ($k_{min} \times A$), turn-on time may not be satisfied due to insufficient drive current. Alternatively, if more area is allocated to sleep transistors than the maximum permitted ($k_{max} \times A$), the constraint on leakage current may not be satisfied since subthreshold leakage current is proportional to device size. This tradeoff has been well characterized in the literature.

In the next step, the proposed analytic expressions (described in the following section) are used to determine the optimum resource allocation to minimize power supply noise. If this optimum allocation is within the permitted range determined in the previous steps, (*i.e.*, $k_{min} \leq k_{opt} \leq k_{max}$), then the number of TSVs and sleep transistor size are finalized, enhancing the system-wide power integrity of the 3D IC, while satisfying turn-on time and leakage current. If the optimum value is outside the range, the minimum power supply noise cannot be obtained under the area constraint A determined in step 1. In this case, the available area may be adjusted to achieve minimum power supply noise, or either k_{min} or k_{max} can be used to allocate the area.

B. Proposed Analytic Expressions to Determine k_{opt}

Assuming a TSV diameter of d, the area of a single TSV is $\pi d^2/4$. The substrate area occupied by a TSV is increased by α due to keep-out zone. Thus, the number of TSVs that can be reliably fabricated within an area of $(1-k) \times A$ is

$$N_{\text{TSV}} = \frac{(1-k)A}{\alpha(\pi d^2/4)}. \tag{1}$$

Similarly, the physical area $k \times A$ consumed by the sleep transistors (each having a channel width W and length L) is estimated as

$$kA = \beta \sum_{1}^{N_{st}} (W \times L), \tag{2}$$

where N_{st} is the overall number of sleep transistors and the parameter β considers the area overhead of the transistors due to drain/source contacts and required spacing between adjacent transistors.

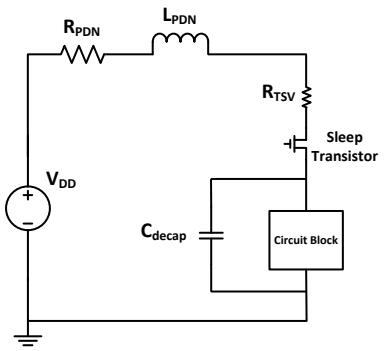

Fig. 2. Simplified model of a power delivery path illustrating sleep transistors and TSVs.

The effective resistance of N TSVs in a bundle is approximated as

$$R_{\text{TSV}}^{\text{eff}} = \frac{4\rho h_{\text{TSV}}}{\pi d^2 N_{\text{TSV}}} = \frac{\rho h_{\text{TSV}} \alpha}{(1-k)A}, \tag{3}$$

where h_{TSV} is the height of a single TSV and ρ is the resistivity of the TSV filling material.

When turned on, the sleep transistors operate in the linear region due to small voltage drop across the source and drain terminals. The channel resistance in this region can be estimated as

$$R_{st} \approx \frac{L}{\mu C_{ox} W (V_{gs}^{st} - V_{th})}, \tag{4}$$

where μ is the electron (or hole) mobility, V_{gs}^{st} is the control signal applied to the gate of the sleep transistor, and V_{th} is the threshold voltage. Note that the sleep transistors typically have high threshold voltages to reduce leakage current consumption during standby mode. Combining (2) and (4), the effective resistance of the sleep transistor is

$$R_{\text{st}}^{\text{eff}} \approx \frac{L^2 \beta}{\mu C_{ox}(V_{gs}^{st} - V_{th})} \times \frac{1}{kA}. \tag{5}$$

The simplified model shown in Fig. 2 is used to gain an intuitive understanding of the proposed methodology. This model describes the power delivery path for a single plane in a 3D stack with distributed power gating topology. The resistance of the TSVs delivering power to this plane is modeled by R_{TSV}. The sleep transistor is placed between the TSV and circuit blocks in the power delivery path. All of the other impedances including the package level parasitics and impedances of the neighboring planes are represented by R_{PDN} and L_{PDN}. The decoupling capacitance of the plane is modeled by C_{decap}.

When evaluating a power distribution network, considering only the static IR drop is not sufficient due to resonance [12]. The method described in [13] is adopted in this paper, which simultaneously considers the static IR drop and resonant supply noise. According to the model

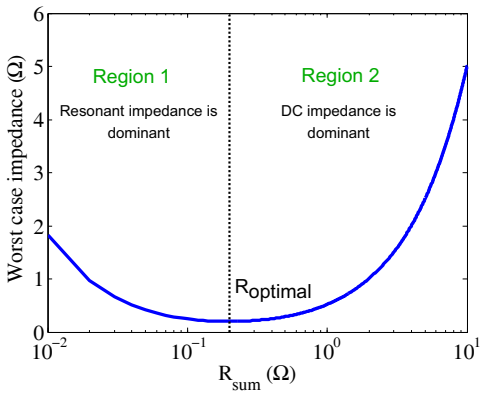

Fig. 3. The magnitude of worst case impedance Z_{worst} varying with R_{sum} values.

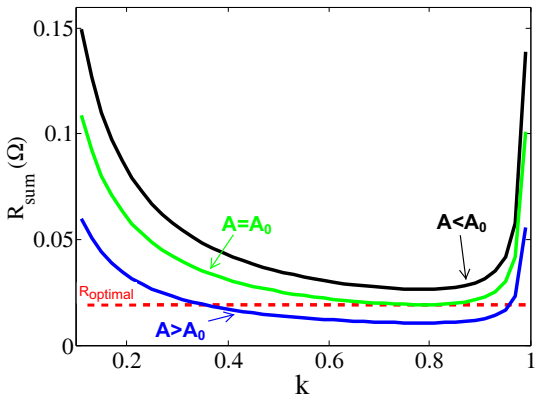

Fig. 4. The R_{sum} under different k values with three area constraint A = 11000 μm^2, A = 8000 μm^2 and A = 20000 μm^2.

in Fig. 2, the impedance of the power supply network is

$$Z(\omega) = (R_{\rm PDN} + j\omega L_{\rm PDN} + R_{TSV} + R_{ST}) \parallel \frac{1}{j\omega C_{decap}}. \quad (6)$$

Thus, the impedance at DC can be expressed as

$$Z_{DC} = R_{\rm PDN} + R_{TSV} + R_{ST}, \quad (7)$$

whereas the magnitude of the impedance at the resonant frequency can be approximated as

$$Z_{res} \approx \frac{L_{\rm PDN}}{(R_{\rm PDN} + R_{TSV} + R_{ST})C_{decap}}. \quad (8)$$

According to [13], the worst case power supply noise is determined by the sum of DC noise and resonant noise,

$$V_{noise}^{worstcase} = V_{noise(DC)} + V_{noise(res)} = Z_{DC} \cdot I_{dc} + Z_{res} \cdot I_{res}, \quad (9)$$

where I_{dc} and I_{res} are, respectively, the current flow at DC and resonant frequency. Assume that the ratio between I_{dc} and I_{res} is μ/ν, where $\mu+\nu = 1$. Therefore, the worst case impedance of the power network is

$$Z_{worst} = \frac{V_{noise}}{I_{dc} + I_{res}} = \mu Z_{DC} + \nu Z_{res}. \quad (10)$$

Z_{worst} is used as a metric to evaluate the power distribution network where static IR drop and resonant noise are both considered. Using (7) and (8) in (10),

$$Z_{worst} \approx \mu(R_{\rm PDN} + R_{TSV} + R_{ST}) + \frac{\nu L_{\rm PDN}}{(R_{\rm PDN} + R_{TSV} + R_{ST})C_{decap}}. \quad (11)$$

$R_{\rm ST} + R_{TSV}$ is defined as R_{sum} and Z_{worst} is plotted as a function of R_{sum} in Fig. 3. As demonstrated in this figure, if R_{sum} is relatively small (region 1), Z_{worst} decreases as R_{sum} increases since the resonant impedance is more dominant than the DC impedance in this region. A larger R_{sum} reduces the resonant impedance (due to greater damping), which decreases Z_{worst}. As R_{sum} further increases, Z_{worst} reaches a minimum and starts increasing (in region 2). The increase in region 2 is due to the DC impedance, which is the dominant term in this region. Z_{worst} is minimized when the two terms on the right-hand-side of (11) are equal. This condition is satisfied when

$$R_{sum} = R_{optimal} = \sqrt{\frac{\nu}{\mu} \cdot \frac{L_{\rm PDN}}{C_{decap}}} - R_{ST}. \quad (12)$$

Thus, k should be chosen such that R_{sum} is equal to the $R_{optimal}$. Using (3) and (5), R_{sum} is be expressed in terms of k as

$$R_{sum} = \frac{T_1}{kA} + \frac{T_2}{(1-k)A}, \quad (13)$$

where

$$T_1 = \frac{L^2\beta}{\mu C_{ox}(V_{gs}^{st} - V_{th})}, \quad T_2 = \rho h_{\rm TSV}\alpha.$$

Replacing (12) in (13),

$$\frac{T_1}{kA} + \frac{T_2}{(1-k)A} = \sqrt{\frac{\nu}{\mu} \cdot \frac{L_{\rm PDN}}{C_{decap}}} - R_{ST}. \quad (14)$$

The two real roots of (14), given by (15), achieve $R_{optimal}$, thereby minimizing the worst case impedance of the power network, Z_{worst}.

C. Area Dependence

It is important to note that $R_{optimal}$, determined by (12), is independent of the physical area available to sleep transistors and TSVs. Thus, if the overall area allocated to sleep transistors and TSVs is smaller than a certain value, $R_{optimal}$ cannot be achieved by any k value. To illustrate this phenomenon, R_{sum} is plotted in Fig. 4 as a function of k under three different area constraints. As shown in this figure, if the available area is smaller than the critical area, indicated as A_0, the R_{sum} curve does not reach $R_{optimal}$ for any value of k. Alternatively, if the area is larger or equal to A_0, then $R_{optimal}$ is achieved at a particular k, as

$$k_{optimal1,2} = \frac{(R_{optimal}A + T_1 - T_2) \pm \sqrt{(R_{optimal}A + T_1 - T_2)^2 - 4R_{optimal} \cdot AT_1}}{2R_{optimal} \cdot A} \quad (15)$$

analytically determined by (15). Note that for the R_{sum} curve when $A > A_0$, there are two k values where $R_{optimum}$ is achieved. According to the proposed flow, as shown in Fig. 1, the optimum k should be between k_{min} and k_{max} to satisfy the constraints on turn-on time and leakage current. Thus, the selected k should be within this range. Note that a larger k favors turn-on time due to wider sleep transistors whereas a smaller k favors leakage current due to smaller sleep transistors.

The critical area A_0 can also be analytically expressed. According to Fig. 4, the minimum value of R_{sum} is equal to $R_{optimal}$ if $A = A_0$. The k value that achieves minimum R_{sum} is

$$\frac{dR_{sum}}{dk} = 0 \rightarrow k_{min} = \frac{1}{\sqrt{\frac{T_2}{T_1}} + 1}. \quad (16)$$

If this k_{min} is used, the minimum R_{sum} is obtained as,

$$R_{sum}^{min} = \frac{(\sqrt{T_1} + \sqrt{T_2})^2}{A}. \quad (17)$$

Since $R_{sum}^{min} = R_{optimal}$ at $A = A_0$, A_0 can be determined by equating (12) with (17), and solving for A,

$$A_0 = \frac{(\sqrt{T_1} + \sqrt{T_2})^2}{(\sqrt{\frac{\nu}{\mu} \cdot \frac{L_{\text{PDN}}}{C_{decap}}} - R_{ST})}. \quad (18)$$

Thus, if the overall area allocated to TSVs and sleep transistors is equal or greater than (18), the worst case impedance, as determined by (11), can be minimized.

III. Simulation Results

To evaluate the proposed methodology and analytic expressions, a comprehensive simulation setup is developed, as described in Section III-A. The results are presented in Section III-B.

A. Simulation Setup

To evaluate the proposed methodology and analytic expressions, an evaluation setup for a three-plane 3D IC is developed, as illustrated in Fig. 5. A 45 nm CMOS technology with 10 metal layers is assumed [14]. The supply voltage V_{DD} is 1 V. A portion of the power network with an area of 1 mm × 1 mm is analyzed using HSPICE. The package level parasitic impedances are modeled with lumped resistance R_{pkg} = 1 mΩ and inductance L_{pkg} = 120 pH.

On each plane, the top two metal layers (Metal 10 & 9) are dedicated to global power grid whereas metal layers 8 and 7 are used as the virtual power grid connected to the global grid through sleep transistors. Both grids use an interdigitated structure with the physical characteristics as listed in Table I. Power gating is achieved using a distributed topology where sleep transistors that control a plane are placed within that plane [10].

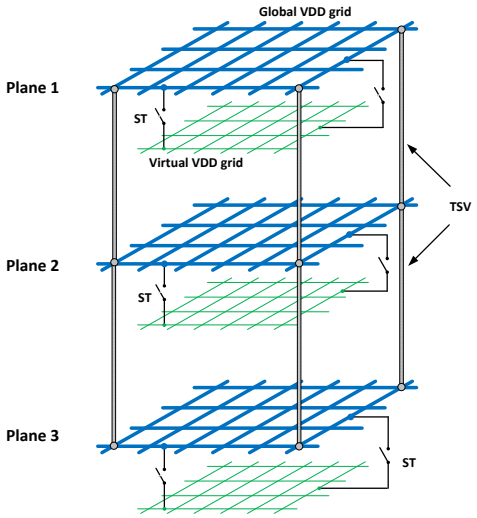

Fig. 5. Power distribution network of a three-plane 3D IC with via-last TSVs and power gating illustrating the global and virtual power grids, sleep transistors (ST), and TSVs.

TABLE I
Primary physical characteristics of the global and virtual power grids [14].

Parameters		Values
Metal 10 & 9	Pitch	45 μm
	Width	40 μm
	Resistivity (ohm/sq)	0.03
Metal 8 & 7	Pitch	23.5 μm
	Width	20 μm
	Resistivity (ohm/sq)	0.075

Via-last TSVs filled with copper are used to interconnect the neighboring planes. The parasitic impedance of a single TSV are R_{tsv}=20 mΩ, C_{tsv} = 283 fF and L_{tsv} =35 pH. Clustered power TSVs are distributed throughout the area as a 5×5 regular array.

To consider the spatial distribution of switching circuits, the overall area on each plane is divided into 30 sub-blocks. For each block, similar to [13], two current sources are used to represent the circuit loads, *i.e.*, a DC current source to produce the I_{dc} and an AC current source at the resonant frequency to produce the I_{res}. The DC and AC current values are both equal to 0.5×I_{peak}, *i.e.*, $\mu = \nu = 0.5$. The peak current I_{peak} of each block is determined based on the power density distribution provided in [15]. Decoupling capacitors are inserted at multiple locations depending upon the power noise distribution.

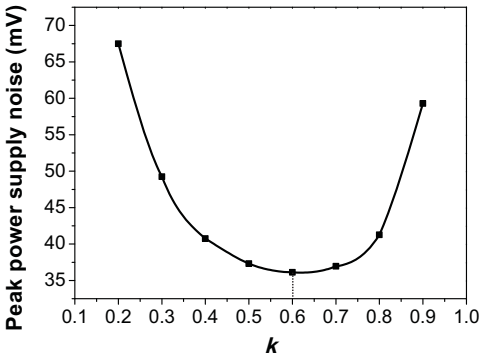

Fig. 6. Simulated peak transient power supply noise of a circuit load in the bottommost plane under different values of k.

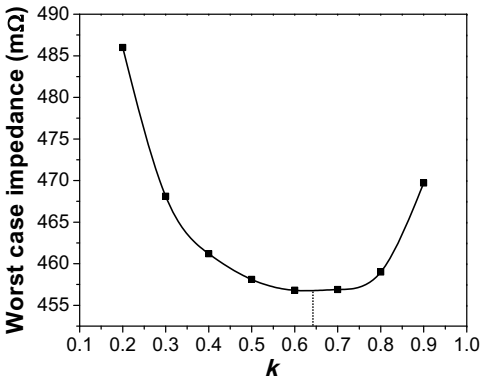

Fig. 7. Worst case impedance for a circuit load in the bottommost plane predicted from the proposed analytical model.

B. Verification of the Proposed Methodology

According to Fig. 1, the first step is to determine the minimum and maximum values of k to ensure that k_{opt} is within this range. The simulation setup described in the previous section is analyzed using HSPICE. k_{min} is determined as 0.06 to guarantee that leakage current is reduced by three orders of magnitude, similar to [16]. Note that higher reduction can be achieved by increasing k_{min}. Similarly, k_{max} is determined as 0.97 to guarantee that the worst case turn-on time for a plane does not exceed 400 ps. A shorter turn-on time can be achieved by decreasing k_{max}.

In the next step, the simulation setup is analyzed to experimentally determine k_{opt}. The worst case power supply noise on the bottom plane is shown in Fig. 6 as a function of k. According to this figure, peak power supply noise is minimized (36 mV) if k is approximately 0.6, *i.e.*, 60% of the overall area is allocated to sleep transistors and the remaining area is used for TSVs. Note that a nonoptimal k can significantly increase the power supply noise. For example, at $k = 0.2$, power supply noise is 67 mV, beyond the 5% constraint.

In the last step, the proposed analytic expressions are utilized to determine k_{opt}. AC simulations are performed to extract the effective impedances required by the proposed model. The calculated worst case impedance (that is correlated with the peak power supply noise) is shown in Fig. 7 as a function of k. As shown in this figure, according to the proposed expressions, k_{opt} is approximately equal to 0.64. The estimated value is sufficiently close to the simulations where the error is 4%. The error is due to the approximations made to extract the effective impedances from the highly distributed simulation setup.

IV. Conclusions

3D ICs are expected to be heavily power gated. A methodology and analytic expressions have been proposed to appropriately allocate available physical area between sleep transistors and TSVs. The methodology minimizes power supply noise while simultaneously satisfying leakage current and turn-on time. The proposed expressions have been verified through a comprehensive simulation setup where the error is less than 4%. The proposed resource allocation achieves more than 46% reduction in supply noise.

References

[1] E. Salman and E. G. Friedman, *High Performance Integrated Circuit Design*. McGraw-Hill, 2012.

[2] V. F. Pavlidis and E. G. Friedman, *Three-Dimensional Integrated Circuit Design*. Morgan Kaufmann, 2009.

[3] S. M. Satheesh and E. Salman, "Power Distribution in TSV-Based 3-D Processor-Memory Stacks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 4, pp. 692–703, December 2012.

[4] D. H. Kim *et al.*, "3d-maps: 3d massively parallel processor with stacked memory," in *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, February 2012, pp. 188–190.

[5] R. Jotwani *et al.*, "An x86-64 Core in 32 nm SOI CMOS," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 162–172, January 2011.

[6] G. Huang *et al.*, "Power Delivery for 3D Chip Stacks: Physical Modeling and Design Implication," in *Proc. of IEEE Conference on Electrical Performance of Electronic Packaging*, 2007, pp. 205–208.

[7] M. Healy and S.-K. Lim, "Distributed TSV Topology for 3-D Power-Supply Networks," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 11, pp. 2066–2079, 2012.

[8] P. Zhou, K. Sridharan, and S. Sapatnekar, "Optimizing Decoupling Capacitors in 3D Circuits for Power Grid Integrity," *IEEE Design Test of Computers*, vol. 26, no. 5, pp. 15–25, September 2009.

[9] K. Kim *et al.*, "Effects of On-chip Decoupling Capacitors and Silicon Substrate on Power Distribution Networks in TSV-based 3D-ICs," in *Proceedings of the IEEE Electronic Components and Technology Conference*, 2012, pp. 690–697.

[10] H. Wang and E. Salman, "Power Gating Methodologies in TSV Based 3D Integrated Circuits," in *Proceedings of the ACM/IEEE Great Lakes Symposium on VLSI*, May 2013, pp. 327–328.

[11] A. Todri *et al.*, "A Study of Tapered 3-D TSVs for Power and Thermal Integrity," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 2, pp. 306–319, February 2013.

[12] E. Salman, E. G. Friedman, R. M. Secareanu, and O. L. Hartin, "Worst Case Power/Ground Noise Estimation Using an Equivalent Transition Time for Resonance,," *IEEE Transactions on Circuits and Systems-I: Regular Papers*, vol. 56, no. 5, pp. 997–1004, May 2009.

[13] J. Gu, H. Eom, J. Keane, and C. Kim, "Sleep Transistor Sizing and Adaptive Control for Supply Noise Minimization Considering Resonance," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 9, pp. 1203–1211, September 2009.

[14] "FreePDK45." [Online]. Available: http://www.eda.ncsu.edu/wiki /FreePDK45:Contents

[15] Q. Zhu, *Power Distribution Network Design for VLSI*. Wiley, 2004.

[16] S. Mutoh *et al.*, "1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 8, pp. 847–854, August 1995.