Thermomechanical Stress-Aware Management for 3D IC Designs

Qiaosha Zou * Tao Zhang * Eren Kursun † Yuan Xie *‡
* The Pennsylvania State University, USA
† IBM Research, Yorktown Heights, USA
‡ AMD Research, Advanced Micro Devices, Inc., USA
‡ Email: *{qszou, taozhang}@cse.psu.edu, † ekursun@us.ibm.com, ‡ yuan.xie@amd.com

***Abstract*—The thermomechanical stress has been considered as one of the most challenging problems in three-dimensional integration circuits (3D ICs), due to the thermal expansion coefficient mismatch between the through-silicon vias (TSVs) and silicon substrate, and the presence of elevated thermal gradients. To address the stress issue, we propose a thorough solution that combines design-time and run-time techniques for the relief of thermomechanical stress and the associated reliability issues. A sophisticated TSV stress-aware floorplan policy is proposed to minimize the possibility of wafer cracking and interfacial delamination. In addition, the run-time thermal management scheme effectively eliminates large thermal gradients between layers. The experimental results show that the reliability of 3D design can be significantly improved due to the reduced TSV thermal load and the elimination of mechanical damaging thermal cycling pattern.**[1]

I. Introduction

Three-dimensional integrated circuits (3D ICs) have been shown to provide number of advantages in achieving higher performance [1]–[3]. Besides the increased interconnectivity provided by the through-silicon vias (TSVs), 3D integration also enables higher packaging density and heterogeneous integration. However, the increased power density induces thermal related issues that have been observed as major barriers in 3D IC designs [4]–[6].

In addition, the Coefficient of Thermal Expansion (CTE) mismatch and thermal expansion directions can adversely affect TSV performance or even crack the entire die, which exacerbates the reliability of 3D ICs. To this end, 3D IC designs should carefully take into account the aforementioned thermal challenges in both design-time and run-time. Unfortunately, little work has been done to alleviate the challenge through these two stages. As a result, in this study we propose a two-stage, design- and run-time solution to this problem. *To our best knowledge, this is the first study that provides dynamic CTE-aware thermal management and cracking prevention in 3D stacked architectures.* The raised two-stage thermal management methodology can be summarized as:

- ***Design-time thermal stress-aware TSV floorplan*** *to reduce the TSV thermal load and minimize thermomechanical stresses to neighbor devices.*
- ***Run-time thermal management scheme*** *to analyze thermal cycling patterns and control thermal expansion and thus to lower the mechanical stress on chip and eventually achieve mechanical equilibrium among each layer.*

The rest of paper is organized as follows. The background of TSV thermal expansion effects and thermal characteristics analysis are introduced in Section II, accompanied with quick review of related work. Proposed design-time and run-time thermomechanical stress-aware floorplan and management methodology are shown in Section III. Block-level and system-level case studies and results analysis are performed in Section IV, followed by Section V to conclude the paper.

[1]This work is supported in part by SRC grants, NSF 0903432, and 1017277.

978-3-9815370-0-0/DATE13/© 2013 EDAA

II. Background and Related Work

Stacked chips on 3D architecture increase the packaging density and thermal resistances, which results in higher on-chip temperatures. Plenty of studies have been done by focusing on the 3D thermal modeling, analysis [7]–[9], and thermal-aware design methodology [10], [11] to manage the on-chip thermal issues of 3D ICs. These work, however, failed to consider the TSV lateral thermal blockage effect and thermomechanical stress. Moreover, they used TSVs as thermal vias to build vertical heat dissipation path, which in turn results in increased thermal load on TSVs as well as thermomechanical stresses, and thus weakens the reliability. On the other hand, prior work on analyzing the mechanical stresses in 3D ICs [12], [13] only consider the static stress management by adjusting TSV keep-out zone size, TSV placement, or TSV structure. Distinguished from previous work, this work not only accounts for the static (design-time) management of TSV thermal stress and thermal load but also takes into account the run-time TSV stress analysis and management.

A. Analysis of TSV thermal stress

In 3D IC fabrication, copper (Cu) is usually used as TSV filling material that has more than **five** times larger CTE than silicon. The CTE mismatch between TSV and silicon substrate in turn introduces mechanical stresses that can lead to high probability of die cracking and interfacial delamination [14], [15].

To minimize the thermomechanical stresses, TSV farms should be placed smartly during design time. Therefore, the corresponding analysis on the thermal stresses around TSVs is critical to the solution. There have been several work [16] [17] targeting on the thermal stresses analysis showing that the stress field in TSVs is uniform and can be represented by radial, circumferential, and axial stresses. The stresses can be expressed as followings:

$$\sigma_r = \sigma_\theta = \frac{-E(\alpha_{tsv} - \alpha_{si})T_{tsv}}{2 - 2\upsilon}, \sigma_z = 2\sigma_\theta \quad (1)$$

where σ_r, σ_θ, and σ_z are radial, circumferential, and axial stresses, respectively. α_{tsv} is the CTE of TSVs and α_{si} represents the CTE of silicon. T_{tsv} is the thermal load on TSV, E is the Young's modulus and υ is the Poisson's ratio[2].

From the equation above, the stresses in TSV are proportional to the thermal load and CTE mismatch. When the material and diameter of TSVs are determined, the only variable is T_{tsv}, the thermal load of TSVs. Therefore, the proposed design-time thermal management scheme alleviates thermomechanical stresses by reducing the thermal load on TSVs.

[2]In this formula, the difference of elastic between materials is omitted for simplicity.

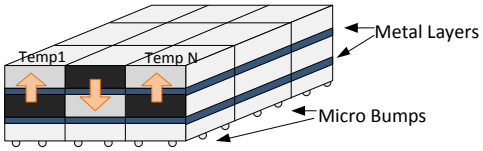

Fig. 1. Stack level thermal cycling effect in 3D structure. Thermal stresses are pointing from hot blocks(dark color) to cool blocks(light color). Alternating direction of stresses (the arrows) easily cause cracking on thinned substrate.

TSVs thermal load estimation during design-time is usually based on accurate thermal modeling of TSVs and TSV temperature is used to represent the corresponding thermal load. Both vertical high thermal conductivity and lateral thermal blockage effect [9] are considered in our TSV model for more accurate temperature modeling.

B. 3D Thermal Cycling Effects

Thermal cycling effect is an other factor that can cause reliability issues of 3D ICs [18]. As shown in Fig. 1, the generated thermal expansion forces are highlighted by arrows, which is from the hotter blocks (dark color) to the cooler blocks (light color). When the force direction varies in the stacked chips, it makes the thinned silicon substrate more vulnerable to be damaged. A run-time thermal cycling management scheme is proposed to eliminate the damaging thermal cycling pattern by using dynamic power scaling.

III. Thermomechanical Stress-aware 3D Design Methodology

In this section, we present the detailed explanation of our proposed design-time thermal stress-aware floorplan technique and run-time thermal management scheme to achieve mechanical equilibrium thermal cycling pattern.

A. The Heuristic Floorplan Flow

The purpose of the design-time thermal stress-aware TSV management aims at reducing the CTE-induced thermomechanical stress. Based on Eq. 1, the reduction of thermomechanical stress can be translated into the minimization of the TSVs thermal load.

Intuitively, placing TSV farms far away from the hot regions can reduce their thermal load. However, in most cases, the hotspots are the functional units that are most active and highly utilized, which usually indicates the requirement for the high connectivity. Moving TSV farms away from those hotspots induces larger wire length and communication delay. As a result, the thermal stress-aware floorplan is obligatory to sustain the high circuit performance without causing severe thermal reliability problems.

In addition to the traditional floorplan solutions that make great efforts on balancing the area and performance trade-off, the novel floorplan flow also strives to minimize the thermal-induced stresses at the same time. Circuit description and average power consumption of each block are given as inputs. The average power consumption of each block is estimated from the assumed power density on the chip. The circuit description consists of: 1) block descriptions, including the block name, area, and allowable aspect ratio (minimum and maximum aspect ratio during floorplanning); and 2) the connectivity information. TSV farms are treated as soft blocks in the floorplan with given thermal characteristics described in Section II.

A simulated annealing based floorplanner is employed in the flow along with an analytical initial floorplan to speed up the convergence. The circuit is partitioned into required tiers by balancing the TSV number and chip area. Then, the initial floorplan is performed analytically by placing the modules that have low power density around TSV farm. Afterwards, the TSV and tier temperature are generated for initial cost calculation. After the initial floorplan, modules are randomly selected and permuted to obtain a better thermal distribution across the whole chip. Besides changing module position, the aspect ratio of modules can be adjusted. For TSV, changing the aspect ratio of TSV farm means adjusting the arrangement of fixed number TSVs. By adjusting the arrangement of TSVs, the lateral thermal path is changed for better heat dissipation.

A cost function is developed as the criteria to evaluate the floorplan result in each simulated annealing iteration 2, where α, β, γ, and δ are associated weighting parameters. A is the final chip area, T_{avg} is the average temperature among blocks, T_{tsv} is the estimated average temperature of TSV farms, and W is the wire length in the design. A simple manhattan-distance based wire length model is used to estimated the wire length overhead.

$$Cost = \alpha * A + \beta * T_{avg} + \gamma * T_{tsv} + \delta * W \quad (2)$$

After each iteration, the cost is calculated based on the floorplan thermal profile to guide the floorplan. The whole iteration terminates once the SA convergence condition is satisfied or maximum iteration step is reached.

B. Thermal Cycling-aware Run-time Management

Run-time thermal cycling management scheme is devised as the second stage of our thermal management methodology to achieve 3D architecture mechanical equilibrium by eliminating mechanical damaging cycling patterns as illustrated in Section II.

The run-time management is performed following a bottom-up, layer-by-layer order for the whole stack. Each tier is partitioned into many small grids, and the temperature of each grid is monitored in runtime. Given the power trace derived from the supply voltage and activity factor of each block, a dynamic thermal profile can be obtained from temperature sensors in each sampling cycle. The proposed dynamic thermal management framework is shown in Fig. 2. After the preliminary sampling period, the grid temperature on each tier is available and the temperature gradients can be captured[3].

The first step in the flow controls the temperature gradients of each grid to eliminate large temperature gradients between adjacent grids. Correspondingly, mechanical force vectors are generated based on this temperature gradients information. On the other hand, The predefined thresholds are determined based on the TSV size, material, and substrate thickness. After the comparison between the force vectors and the predefined thresholds, if the force vectors are larger than thresholds, dynamic power scaling techniques, such as DVFS, will be deployed to control the thermal dissipation of hot grids.

In addition to the temperature gradients, the thermal cycling pattern should be handled carefully. After the control of temperature gradients, the cycling pattern is taken into account by comparing the force vectors of neighboring grids in two adjacent layers. If the thermal cycling pattern is in an alternating way as described in Fig. 1, dynamic power management is applied to the high temperature regions to lower the resulting thermal mechanical stresses and achieve mechanical equilibrium. The power scaling results in new thermal cycling pattern in the stack, which produces further adjustment in next sampling interval till the whole stack reaches the mechanical equilibrium state.

IV. Experiment Results and Analysis

To evaluate the effect of the proposed thermal management methodology, a typical 3D floorplaner 3DFP [19] that aims at

[3]In this work, the temperature differences between grids are used to represent temperature gradients for simplicity.

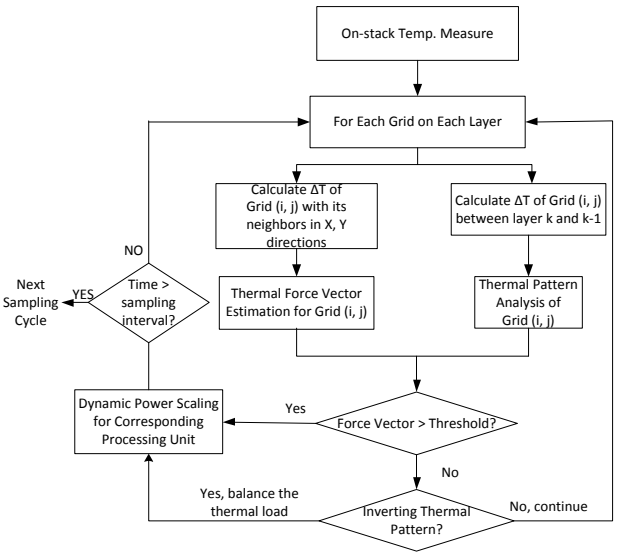

Fig. 2. Run-time Thermal Cycling-aware Thermal Management Flow for One Sampling Cycle.

reducing the average on-chip temperature is employed as our baseline. We successfully implemented the thermal stress-aware floorplan in 3DFP. Hotspot [20] with 3D capability is adapted for circuit temperature evaluation after final floorplan. MCNC benchmarks and a core+memory 3D stacking architecture are leveraged in the block-level and system-level thermal stress-aware floorplan. We differentiate the block-level and system-level simulation since they have distinct thermal characteristics. The temperature gradients at block level are larger and more unpredictable due to the various characteristics of the underlying circuits, while the temperature distribution at system level is usually uniform and predictable because of the relatively regular placement of functional modules.

The TSV farm lateral thermal conductivity ($180W/(mK)$) is derived from the average value reported from previous study [21]. For simplicity, the elastic mismatch between silicon and copper is neglected, so the Young's modulus ($120GPa$) and Poisson's ratio (0.30) are the average values of copper and silicon.

A. Block-level Thermomechanical Stress-aware Floorplan

To quantify the TSV thermal load reduction from the stress-aware floorplan, we conduct experiments on four MCNC benchmarks: `ami33`, `ami49`, `hp`, and `xerox`. Similar to the approach in [19], these benchmarks are partitioned into two tiers at block level. The characteristics of benchmarks vary in terms of block numbers, interconnect complexity and power density. These circuit information is extracted from the benchmark description files. The average power density for each circuit is within the rage of 0.5-2.4 W/mm^2. TSV farms are created based on the partitioning and interconnects information. During the simulation, all modules including TSV farms are treated as soft blocks with flexible aspect ratio. In addition, the weighting parameters for the average on-chip temperature and TSVs thermal load are the same in the cost function.

The experiment results are shown in Table I. As shown, the average TSV temperature reduction is 15.5K and the peak temperature has been reduced after considering the TSV lateral thermal conduction. The chip area for three circuits are slightly reduced due to the re-shape of modules. The execution time for both proposed flow and baseline and TSV axial thermal stress reduction are listed in the last two columns, where the thermal axial stress reduction is 39.01MPa.

Among these four circuits, `ami33` has large number of blocks but the smallest die area. As a result, the average assigned power density is higher than other three benchmarks, which induces highest average and peak temperature. By moving the TSV farm away from hotspots, the TSV temperature can be significantly reduced with slightly area overhead. `ami49` has the lowest average temperature due to the largest chip area and block number. The TSV temperature is only reduced by 3.23K in this circuit because of smaller reduction margin.

B. System-level Thermomechanical Stress-aware Floorplan

In addition to the block-level simulation, we also applied the thermal management flow into a TSV-based 3D design for the system-level analysis. Different from block-level partitioning, the whole system has even higher connectivity requirement, indicating larger TSV occupancy in the floorplan. The 3D design stacks two tiers that have same size in a fashion of face-to-back bonding. The bottom layer mimics a multi-core processor, where four SPARC-like cores with private L2 cache are deployed. The top layer is assumed as a stacked memory chip that is divided into four blocks as the last level cache. TSV buses are integrated as inter-layer connection.

In the baseline, wire length and average temperature are the major metrics to guide the designs. In this way, TSV buses are placed in the middle of the chip to reduce routing length. Fig. 3(a) illustrates the related baseline floorplan. The corresponding zoom-in temperature map is shown in Fig. 4(a). It is obvious that TSV bus has high thermal load since it is surrounded around local hotspots. Due to the lateral thermal blockage effect of TSV bus [9], the heat dissipation path of local hotspots is blocked by TSVs, resulting in elevated temperature and steep temperature gradient.

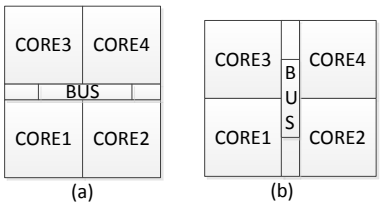

Fig. 3. Floorplan of the core-layer before and after optimization. (a) TSV bus is placed horizontally in the floorplan for wire length and area driven floorplan; (b) TSV bus is placed vertically after thermomechanical stress aware floorplan.

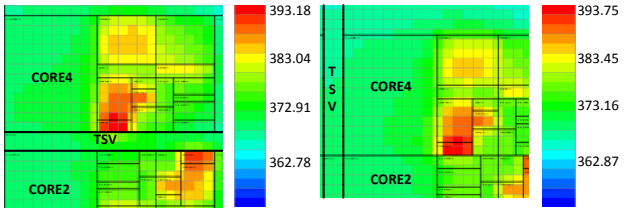

Fig. 4. Zoomed in TSV thermal stress-aware floorplan on core-layer. (a) TSV bus is placed horizontally between execution units of two cores, resulting in higher TSV temperature; (b) TSV bus is placed vertically to for reduced thermal load.

In order to reduce the thermal load and thermal stresses of the TSV bus, the proposed thermal stresses management flow gives another solution after TSVs floorplan optimization as shown in Fig. 3(b). The corresponding temperature map is shown in Fig. 4(b). By placing the TSV bus away from local hotspots, the overall temperature on TSV bus is decreased. In general, the stress-aware floorplan can decreases the average temperature on TSV bus by 6.53K with 16.46MPa axial thermal stress reduction. The peak temperature of each core slightly

TABLE I
DESIGN TIME THERMOMECHANICAL STRESS-AWARE FLOORPLAN RESULTS WITH AVERAGE AND PEAK ON-CHIP TEMPERATURE, NORMALIZED AREA/WIRE LENGTH AND THERMAL STRESSES REDUCTION

Circuit	Baseline			Thermomechanical Stress-aware Placement					Run Time (s)	Stress Reduction
	avg T (K)	peak T (K)	TSV T (K)	avg T (K)	peak T (K)	TSV T (K)	area	wire length	proposed/baseline	(MPa)
ami33	361.55	388.38	365.77	363.53	385.46	338.88	1.08	1.004	40.88/31.11	67.76
ami49	329.45	374.12	317.48	329.83	355.2	314.25	0.98	1.007	201.79/198.56	8.14
hp	339.71	370.06	359.38	338.36	377.08	336.2	0.95	1.14	6.58/5.66	58.41
xerox	346.44	397.21	357.02	343.48	372.87	348.4	0.88	1.13	12.36/12.09	21.72

increases about 4.38K because moving TSV bus away results in direct contact of local hotspots. The average temperature for upper layer memory is 369.54K. The relatively lower temperature on the top layer is beneficial from lower power density and uniform power distribution.

C. System-level Run-time Thermal Management Scheme

The run-time thermal management scheme is examined on the optimized 3D chip in last section. In spite its time-consuming temperature evaluation of HotSpot, we leverage accurate thermal information provided by HotSpot in grid granularity for sensor simulation. Furthermore, the transient power trace of each block as well as the activity factor are also included to help evaluate run-time on-chip temperature.

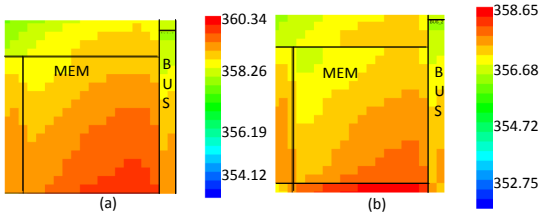

Fig. 5. Zoomed in memory layer run-time thermal management results. The zoomed region causes inverting thermal cycling pattern between two layer. (a) The thermal map before power scaling; (b) thermal map after power scaling on interested region.

As shown in Fig. 5, the transient irregular temperature distribution on the memory layer can cause inverted thermal cycling pattern. Before power scaling, the inverted thermal cycling pattern may occur between the cooler region of lower core layer and the hotter region of upper memory layer. In order to have a better understanding of this thermal cycling effect, each memory block is further partitioned into three regions. The region in the memory (shown in Fig. 5) causes the mechanical damaging thermal cycling pattern. After the power scaling on the corresponding memory region, the temperature of this region is decreased and the mechanical damaging thermal cycling pattern is eliminated. The average temperature on memory chip is reduced by 1.5K after the power scaling, resulting smaller temperature gradients between two layers. In this case study, the power scaling is performed by reducing the frequency from 1333MHz to 1066MHz on memory layer. The performance degradation is highly related to the application characteristics, on average, the performance degrades less than 5% after frequency scaling [22].

V. CONCLUSION

This paper presents a two-stage thermal management technique on design-time and run-time to alleviate the thermal challenges on 3D architectures. The design-time TSV thermal stress-aware floorplan technique aims at reducing the TSV thermal load during floorplan in design time. In run time, thermal gradients and thermal cycling pattern induced mechanical reliability challenges are considered. Controlling the temperature gradients and eliminating damaging thermal cycling patterns can reduce the risk of cracking on the thinned silicon substrate. The results show that design-time floorplan can effectively reduce TSV thermal load for thermomechanical stresses minimization. The radical thermal stress reductions on average are 39.01MPa and 16.46MPa in block-level and system-level case studies, respectively. Experiment results illustrate that after run-time management, the core to memory stacking can achieve mechanical equilibrium on thermal cycling through dynamic power scaling with slightly performance overhead.

REFERENCES

[1] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. Franzon, "Demystifying 3D ICs: The Pros and Cons of Going Vertical," *Design Test of Comoputer, IEEE*, vol. 22, 2005.

[2] G. H. Loh, Y. Xie, and B. Black, "Processor Design in 3D Die-Stacking Technologies," *IEEE MICRO*, 2007.

[3] Y. Xie, G. H. Loh, B. Black, and K. Bernstein, "Design Space Exploration for 3D Architectures," in *J. Emerg. Technol. Comput. Systs.*, 2006.

[4] P. Leduca, F. de Crecy, M. Fayolle, B. Charlet, T. Enot, M. Zussy, B. Jones, J.-C. Barbe, N. Kernevez, N. Sillon, S. Maitrejean, and D. Louisa, "Challenges for 3D IC Integration: Bonding Quality and Thermal Management," in *International Interconnect Technology Conference*, 2007.

[5] K. Puttaswamy and G. H. Loh, "Thermal Analysis of a 3D Die-Stacked High-Performance Microprocessor," in *GLVLSI*, 2006.

[6] S. Das, A. Chandrakasan, and R. Reif, "Timing, Energy, and Thermal Performance of Three-Dimensional Integrated Circuits," in *GLVLSI*, 2004.

[7] A. Jain, R. Jones, R. Chatterjee, and S. Pozder, "Analytical and Numerical Modeling of the Thermal Performance of Three-Dimensional Integrated Circuits," *Components and Packaging Technologies*, 2010.

[8] F. Wang, Z. Zhu, Y. Yang, and N. Wang, "A Thermal Model for the Top Layer of 3D Integrated Circuits Considering Through Silicon Vias," in *ASICON*, 2011.

[9] Y. Chen, E. Kursun, D. Motschman, C. Johnson, and Y. Xie, "Analysis and Mitigation of Lateral Thermal Blockage Effect of Through-Silicon-Via in 3D IC Designs," in *ISLPED*, 2011.

[10] J. Cong, G. Luo, J. Wei, and Y. Zhang, "Thermal-Aware 3D IC Placement Via Transformation," in *ASPDAC*, 2007.

[11] J. Cong, G. Luo, and Y. Shi, "Thermal-Aware Cell and Through-Silicon-Via Co-Placement for 3D ICs," in *DAC*, 2011.

[12] K. Athikulwongse, A. Chakraborty, J.-S. Yang, D. Pan, and S. K. Lim, "Stress-Driven 3D-IC Placement with TSV Keep-Out Zone and Regularity Study," in *ICCAD*, 2010.

[13] M. Jung, J. Mitra, D. Pan, and S. K. Lim, "Tsv Stress-aware Full-Chip Mechanical Reliability Analysis and Optimization for 3D IC," in *DAC*, 2011.

[14] K.-H. Lu, S.-K. Ryu, J. Im, R. Huang, and P. Ho, "Thermomechanical Reliability of Through-Silicon Vias in 3D Interconnects," in *IRPS*, 2011.

[15] C. Selvanayagam, J. Lau, X. Zhang, S. Seah, K. Vaidyanathan, and T. Chai, "Nonlinear Thermal Stress/Strain Analyses of Copper Filled TSV (Through Silicon Via) and Their Flip-Chip Microbumps," *Advanced Packaging*, 2009.

[16] S.-K. Ryu, K.-H. Lu, X. Zhang, J.-H. Im, P. Ho, and R. Huang, "Impact of Near-Surface Thermal Stresses on Interfacial Reliability of Through-Silicon Vias for 3-D Interconnects," *Device and Materials Reliability*, 2011.

[17] K. H. Lu, X. Zhang, S.-K. Ryu, J. Im, R. Huang, and P. Ho, "Thermo-Mechanical Reliability of 3-D ICs Containing Through Silicon Vias," in *ECTC*, 2009.

[18] C. Noritake, P. Limaye, M. Gonzalez, and B. Vandevelde, "Thermal Cycle Reliability of 3D Chip Stacked Package Using Pb-free Solder Bumps: Parameter Study by FEM Analysis," in *EuroSimE*, 2006.

[19] W.-L. Hung, G. Link, Y. Xie, N. Vijaykrishnan, and M. Irwin, "Interconnect and Thermal-Aware Floorplanning for 3D Microprocessors," in *ISQED*, 2006.

[20] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. Stan, "Hotspot: A Compact Thermal Modeling Methodology for Early-Stage VLSI Design," *VLSI*, 2006.

[21] Z. Chen, X. Luo, and S. Liu, "Thermal Analysis of 3D Packaging with a Simplified Thermal Resistance Network Model and Finite Element Simulation," in *ICEPT-HDP*, 2010.

[22] H. David, C. Fallin, E. Gorbatov, U. R. Hanebutte, and O. Mutlu, "Memory Power Management via Dynamic Voltage/Frequency Scaling," in *International Conference on Autonomic Computing*, 2011.