Hierarchical Simulation and Optimization of 3D Power Delivery Network

Song Yao[1], Junior Undergraduate, ACM Student Member, Number 0717121
Advisor: Associate Professor Yu Wang[2]
[1,2]Department of Electronic Engineering, TNList, Tsinghua University, Beijing, China, 100084
[1]yaos11@mails.tsinghua.edu.cn, [2]yu-wang@tsinghua.edu.cn

I. Problem and Motivation

With the feature size shrinking, traditional electronic design methodologies face some bottlenecks such as per wafer cost, large interconnect delay, and high leakage power. Three-dimensional integration circuit (3D IC) has been regarded as a promising solution to mitigate these problems. 3D ICs stack separate tiers in the vertical direction and connect them using vertical connections such as Through-Silicon-Vias (TSVs). Compared with traditional 2D integration, the 3D technology can provide many benefits such as reduction in the interconnect wire length, improvement of memory bandwidth, and support for realization of heterogeneous integration [1].

3D power delivery network (PDN) is the basic support system for 3D ICs, which includes the power/ground (P/G) networks, the vertical interconnects, and the package (shown in Figure 1). In 3D PDNs, P/G TSVs connect the P/G networks in different tiers, and the bottommost tier is connected to the package through metal bumps and then routed to the voltage sources on the PCB board. The design of a 3D PDN is much more complex than the design of a 2D power network [2] and considered as one of the most critical challenges in IC design. As power coupling between different tiers becomes tighter and multiple tiers are stacked together into small footprints, delivering current to all parts of the 3D IC becomes highly challenging. An irrational PDN design may result in low supply voltage levels in the grid, i.e. large IR-drop, and thus increase the gate delay, lead to timing violations and harm the functionality of circuits [3].

To design a PDN, suitable simulation approaches are necessary. However, PDN analysis is usually a time and resource consuming task since there are millions of nodes in most mesh-based PDNs. The simulation problem becomes tougher in 3D cases as the network scale may be several times larger than that of 2D cases. Nevertheless, most existing studies tend to regard the 3D power system as a whole like a 2D power system, which does not utilize the inherent 3D structure and causes inefficiency. Considering that TSVs naturally form good partition boundaries, a hierarchical technique to analyze the 3D PDN regarding the power grid of each layer as a macro-model may lead to efficiency and offer an opportunity for further parallel speedup.

Numerous studies spring up these years in the field of 3D PDN design. However, most studies tended to simply increase the number of TSVs to mitigate power supply problems. Since TSV fabrication causes tensile stress around TSVs and results in significant carrier mobility variation in the devices

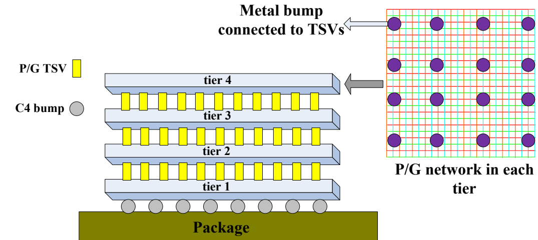

Fig. 1. 3D power delivery network.

near TSVs, additional keep-out zone (KOZ) is adopted to prevent any devices/cells from being impacted. Owing to the relative huge size, large amount KOZ can significantly reduce the available placement area for cells and result in a high cost [4]. To reserve more TSV resources and reduce the overhead for connections, we propose a region-aware P/G TSV planning to optimize the 3D PDN. The proposed method considers a non-evenly TSV topology, a distributed IR-drop constraints, and the whitespace limitation to generate a valid TSV topology which can meet the power supply requirements while reducing the number of P/G TSVs.

II. Background and Related Work

Researchers have proposed many techniques to accelerate P/G analysis for 2D chips such as some multigrid based approaches [5], [6]. The multigrid method can accelerate the convergence of traditional iterative methods through interpolation between fine and coarse grids. A memorized supernodal technique and some parallel forward and back substitution methods were also proposed [7], [8]. Furthermore, Zhao *et al.* presented a hierarchical analysis approach for 2D P/G networks using macro-models [9]. The "divide and conquer" approach in P/G network analysis has a profound effect on researchers of this area.

As for 3D P/G network simulation, some techniques such as the extension of compact physical model [10] were proposed. However, the scalability cannot be ensured with those techniques. Hu *et al.* suggested using standard reduced power models (SRPM) to replace some tiers of 3D P/G network which can also help to mitigate the conflict between data sharing and chip protection [11]. Nevertheless, the key issue is how to build SRPM for each tier. A voltage propagation method for static 3D P/G network analysis was proposed by Zhang *et al.* [12]. However, the selection of the initial voltages has a large impact on the performance of their method.

Fruitful researches have explored the 3D PDN design methodologies. Most of existing studies employed the evenly

distributed TSV topology such as a 3D IC floorplan and P/G network co-synthesis tool [13] and a simultaneous power delivery and thermal integrity optimization method [14]. Increasing the TSV density in the evenly TSV topology is verified to be a good way to mitigate the IR-drop problem [15], [16]. Besides, several studies considering the non-evenly TSV topology were proposed such as a thermal-aware TSV planning algorithm [17] and another non-evenly TSV planning algorithm [18]. Furthermore, Zhong *et al.* pointed out that the TSV assignment came under the influence of the whitespace distribution in a given 3D floorplan and proposed the whitespace insertion (WSI) algorithm to make the whitespace distribution in the given floorplan more reasonable for TSV insertion [19]. The awareness of whitespace endows TSV planning algorithms with practicability.

III. Uniqueness of the Approach

A. Hierarchical transient simulation

Although all tiers are coupled together, the 3D PDN still has an inherent hierarchical structure. In [20], the authors proposed a hierarchical static simulation method named HS3DPG for 3D P/G network analysis. We extend HS3DPG to transient simulation and verified its performance with several industrial benchmarks. The hierarchical transient simulation method separates all tiers, extracts the transient port equivalent model (PEM) for each tier, reestablishes the global network with PEMs, and finally substitutes the global solution to compute voltages at inner nodes.

The hierarchical transient simulation method is not a standalone approach. When the 3D P/G network is separated into different tiers, the PEM extraction of each tier are independent, and thus the PEMs of different tiers can be computed in full parallelism. Besides, existing techniques such as some parallelism techniques [7], [8] can also be adopted to compute the PEM of each tier. Consequently, the proposed method has a potential to enable a hybrid two-level parallelism. In the present version, a sparse solver CHOLMOD [21] based on Cholesky decomposition and back substitution is used in PEM extraction phase.

The transient PEM is used to mask details of the P/G grid in each layer in global simulation. For one tier, ports are defined as the nodes connected to TSVs (TSV clusters). Since one TSV is connected to the metal layer through a C4 bump, which can be regarded as a super node, the port number of each tier is equal to the TSV number. The proposed PEM of one tier can be formulated as follows:

$$I = J * V + S = \begin{bmatrix} \frac{\partial I_1}{\partial V_1} & \cdots & \frac{\partial I_1}{\partial V_M} \\ \vdots & \ddots & \vdots \\ \frac{\partial I_M}{\partial V_1} & \cdots & \frac{\partial I_M}{\partial V_M} \end{bmatrix} \begin{bmatrix} V_1 \\ \vdots \\ V_M \end{bmatrix} + \begin{bmatrix} S_1 \\ \vdots \\ S_M \end{bmatrix}$$

where M is the number of ports, $I \in R^{M\times 1}$ is a current vector flowing into this tier through these ports, $J \in R^{M\times M}$ is a Jacobi matrix reflecting port dependencies, $V \in R^{M\times 1}$ is a voltage vector at ports, and $S \in R^{M\times 1}$ is a current source vector attached to ports in this tier. As shown in Figure 2,

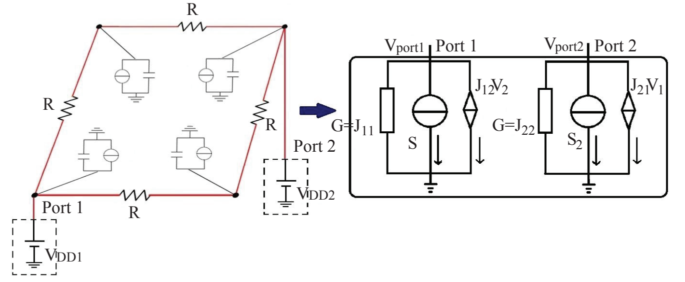

Fig. 2. PEM extraction when port number is 2.

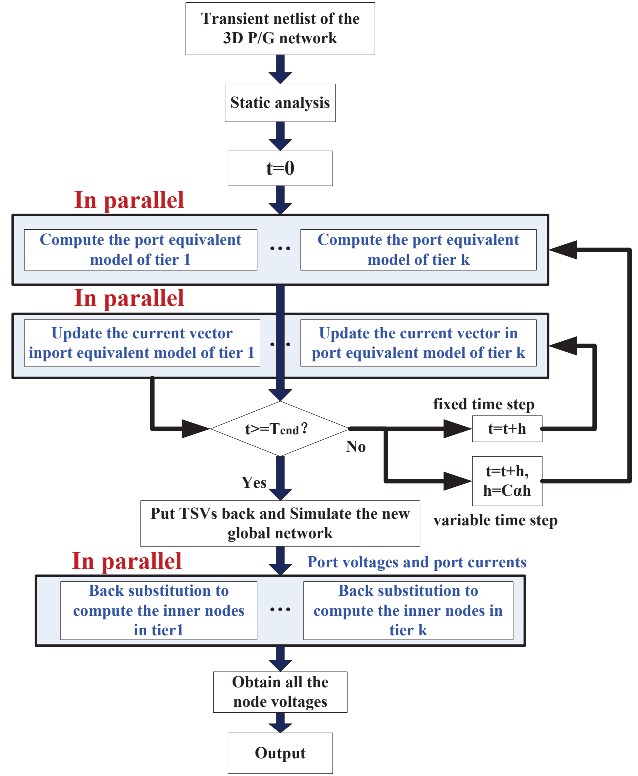

Fig. 3. The hierarchical transient simulation flow.

for circuit presentation, all entries in the Jacobi matrix J are modeled as voltage-controlled current sources (VCCS) and all entries in vector S are modeled as current sources.

In the transient PEM, the entries of the current vector S are modeled as piece-wise linear (PWL) sources. Besides, in the variable-time-step transient simulation, the entries of the Jacobi matrix J are also PWL models. In the fixed-time-step case, as the Jacobi matrix J always remains the same, we only need to update the current vector S in the PEM whose computation complexity is linear (one-time back substitution at every step).

Figure 3 shows the proposed hierarchical transient simulation flow. **(1)When the time step is fixed**, as mentioned above, we only need to update the vector S in transient PEM at every step. In this case, postulating that there are k same tiers in the 3D P/G network and the computation complexity of Cholesky decomposition is $O(N^m)$, the speedup of the hierarchical method can be estimated as:

$$\frac{T_{Direct\ Solve}}{T_{HS3DPG\ Trans}} = \frac{Mk^m + nk}{M + 2n} \quad (1)$$

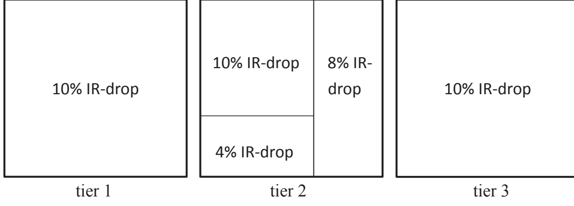

Fig. 4. An example of IR-drop constraints in different tiers.

where n is the number of time steps, $M = \frac{T_{decomposition}}{T_{back\ substitution}}$ is a large constant, $T_{Direct\ Solve}$ is the runtime regarding the 3D P/G network as a whole, and $T_{HS3DPG\ Trans}$ is the runtime using the hierarchical method. Generally, m is about 1.5. As a result, we can find that the hierarchical method is faster than the direct full network simulation as long as $k \geq 2$. On the other hand, the speedup declines with the increase of the number of time steps when k is fixed. This is because in the hierarchical transient simulation, double back substitutions are needed in at every time point. The cost of extra back substitution will accumulate over time and decrease the performance. **(2)When the time step is variable**, the new time step h_{i+1} should be computed follows h_{i+1}=$C\alpha h_i$ at every step. The coefficient C (0.8 $\sim$ 0.9) is a safety parameter and $\alpha = \sqrt{\frac{e_{max}}{|E_{T,i}|}}$, where $e_{max} = \frac{E_{max}}{T} h_i$ and $E_{T,i} = -\frac{h_i^2}{2} x^{''}(t_i)$ are the max allowed error and the local truncation error in the last time step. As the time step h updates, the coefficient matrix $A = G + C/h$ changes at every time step. Consequently, additional Cholesky decompositions are needed in both computing PEMs and substituting to tiers at every time point.

This work was submitted to IEEE Transactions on very large scale integrated (VLSI) systems [22].

B. Region-aware P/G TSV planning

The P/G TSV planning problem for 3D PDN design can be formulated as following: Given the 2D P/G networks in different tiers, the IR-drop constraints of all tiers (shown in Figure 4), and the limitation of available power TSV resources, to generate a TSV distribution topology and output voltage distribution maps.

The evenly TSV planning algorithm is shown in Algorithm 1, which is regarded as the baseline to measure the effectiveness of our region-aware method. In each iteration, we run the static analysis and check the validity. If not all nodes meet the IR-drop constraints, the pitch between TSVs will be reduced, and thus the TSV density can be increased.

The region-aware TSV planning algorithm is shown in Algorithm 2, where $V[i]$ is the voltage at node i, V_{DD} is the ideal supply voltage, $IR\text{-}drop[i]$ is the IR-drop constraint at node i, and *AffectArea* is the estimated area that one TSV can affect. The region-aware method employs a chip segmentation, distribute TSV resources over different areas according to their IR-drop constraints separately while restraining all TSVs in the whitespace. A direct method to determine the node to insert TSV is used: We sweep all the nodes and insert TSVs at nodes whose IR-drop constraints are not met. This method allows inserting numerous TSVs during one iteration,

Algorithm 1: Evenly P/G TSV Planning Algorithm

```
  Input: P/G Networks, IR-drop constraints
  Output: TSV topology, Voltage distribution maps
1 Initial the TSV topology with a single TSV;
2 Executing static P/G analysis;
3 while the IR-drop constraints are not met do
4 |   Increase the TSV density;
5 |   Executing static P/G analysis;
6 end
7 return TSV topology and voltage distribution maps
```

Algorithm 2: Region-Aware P/G TSV Planning Algorithm

```
   Input: P/G Networks, IR-drop Constraints, TSV number
          constraint
   Output: TSV topology, Voltage distribution maps
1  Initial the TSV topology with evenly TSV planning algorithm
   according to the most loose IR-drop constraint;
2  while the TSV number constraint is not violated do
3  |   Execute static P/G analysis;
4  |   for the first node→the last node do
5  |   |   if V[i]<V_DD−IR-drop[i] and node i is not valid then
6  |   |   |   Insert a TSV in the nearest whitespace of node i;
7  |   |   |   Mark the nodes in the TSV's AffectArea as
   |   |   |   valid;
8  |   |   end
9  |   |   Go to the next node;
10 |   end
11 |   if No TSV is added in the latest iteration then
12 |   |   return TSV topology and voltage distribution maps;
13 |   end
14 end
15 return no solution;
```

and thus reduces the number of iterations. Since one iteration requires a static analysis and usually causes quite a few time consumption, the planning process can be accelerated greatly correspondingly with this direct method.

In 3D PDNs, TSVs play the role of C4 bumps in 2D PDNs and can ensure the supply voltage in their neighborhood areas. We define this area as the *AffectArea*. When a TSV is inserted, the nodes in its *AffectArea* are marked as valid. The parameter *AffectArea* can be computed following:

$$AffectArea = \frac{I_{TSV}}{\alpha \cdot \frac{P}{V_{DD}}} \tag{2}$$

where I_{TSV} is current load capacity of one TSV, P is the average power consumption per unit area, V_{DD} is the supply voltage, and α is parameter related to the wire size and the clearance between wires. A suitable parameter *AffectArea* should be used following Equation (2). If the $AffectArea$ is set to be much smaller than the area that one TSV can actually guarantee, redundant TSVs will be inserted in the iterations and the efficiency of the generated TSV topology will degrade. One the contrary, if the $AffectArea$ is too large, only a few TSVs will be inserted during one iteration, and thus more iterations will be needed to generate the suitable TSV topology.

This work was published at ISQED 2014 [23].

TABLE I
RUNTIME COMPARISON IN FIXED-TIME-STEP TRANSIENT SIMULATION

Simulation time (ns)			1	10	20	50	100	500	1000	10000
Testbench name	3D-μp	Direct full simulation	1.617s	3.978s	6.061s	14.006s	25.239s	131.171s	245.136s	2215.605s
		HS3DPG Trans	0.841s	1.963s	4.644s	10.370s	21.553s	109.937s	216.354s	2057.201s
	3D-TxRx	Direct full simulation	0.301s	0.732s	1.199s	3.120s	4.796s	22.547s	44.807s	453.641s
		HS3DPG Trans	0.173s	0.416s	0.701s	2.014s	3.111s	14.338s	28.430s	303.760s

TABLE II
RUNTIME COMPARISONS IN VARIABLE-TIME-STEP TRANSIENT SIMULATION

Simulation time (ns)		1	5	10	15	30	50
3D-μp	Direct	18.234s	92.112s	201.341s	309.798s	421.230s	587.122s
	HS3DPG Trans	11.452s	15.247s	46.186s	60.558s	168.489s	197.174s

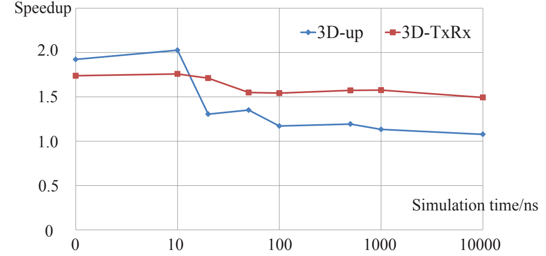

Fig. 5. Speedup curves in the fixed-time-step transient simulation.

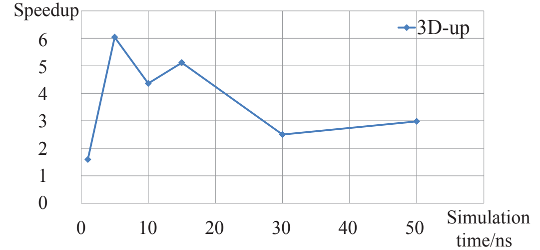

Fig. 6. Speedup curve in the variable-time-step transient simulation.

IV. RESULTS AND CONTRIBUTIONS

A. Hierarchical Transient Simulation

In this section, we verify the performance of the proposed hierarchical transient simulation with benchmarks extracted from industrial designs. The benchmark 3D-μp is a 2-tier multi-processor design and 3D-TxRx is a hardware cryptography design with 3 stacked tiers [24]. We compare the results of the proposed method with the direct full simulation.

In the transient simulation, the complexity of computing PEMs differs greatly in fixed-time-step case and variable-time-step case. However, the proposed method can provide considerable speedup in both cases. **In the fixed-time-step case**, we run the simulation from 1ns to 10μs and the results are shown in Table I and Figure 5. The time step is set to be 0.1ns. Results show the speedup of HS3DPG declines with the increase of simulation time and approaches a limit as we rough estimated. **In the variable-time-step case**, the results are shown in Table II and Figure 6. The max allowed error is set to be 1mV. In this case, the speedup does not degrade with the accumulation of time steps. As the waveforms in benchmark 3D-μp change rapidly, the time step length is compressed, and thus the actual runtime with variable time step is much longer than that of the fixed-time-step case.

The main contributions of this work are listed as follows:

1) We extend the static hierarchical simulation method in [20] to transient simulation and use benchmarks extracted from industrial designs to verify its performance.
2) We explore different forms of PEM and evaluate their performance in fixed-time-step and variable-time-step transient simulation cases. In the fixed-time-step case, the computation of PEM needs two back substitutions at every step and the speedup may decline with the increase of simulation time. In the variable-time-step case, PEM can always lead to a high speedup.

B. Region-aware TSV planning

To evaluate the effectiveness of the proposed region-aware planning method, we compare the TSV consumption with different TSV topologies on several benchmarks. The benchmark 1 consists of 3 tiers and the IR-drop constraints are set according to Figure 4. In this benchmark, the average current density is $128A/cm^2$ and the wire parameters are extracted through PTM model under 45nm technology [25]. Benchmark S-ami33 and S-ami49 are generated following the floorplan results in [17] with under the same technology. We also generate benchmarks to simulate the functional areas distribution in the benchmarks extracted from industrial designs [24]. An example is shown in Figure 7. Since the total number of P/G TSVs is only eight in the original benchmark, which is inconvenient for comparison, we adjust the current density and the P/G wire parameters through the PTM model. The IR-drop constraints are set according to the power consumption maps. The whitespace limitation is not considered on benchmark 1, S-ami33, and S-ami49.

All the results are shown in Table III. Without whitespace limitation, the region-aware planning method can lead to a 41.8 % TSV consumption reduction on average. Considering the whitespace limitation, the region-aware method can also save 27% P/G TSVs. The chip area and the cost can be saved accordingly. We also draw full static voltage distribution map of benchmark 1 and S-3D-μP in Figure 9 and Figure 8 to

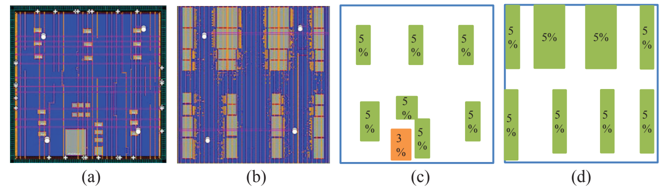

Fig. 7. (a),(b) tier1 and tier 2 of benchmark 3D-μp [24]; (c),(d) the inputs for the region-aware method to simulate the benchmark.

TABLE III
RESULTS OF TSV CONSUMPTION WITH DIFFERENT TSV PLANNING METHODS

Benchmark	Tiers	TSV(Region-Aware)	TSV(Evenly)	TSV reduction
benchmark 1	3	84	162	48.1%
S-ami33	4	187	300	37.7%
S-ami49	4	147	243	39.5%
S-3D-μp	2	76	100	24.0%
S-3D-TxRx	3	140	200	30.0%

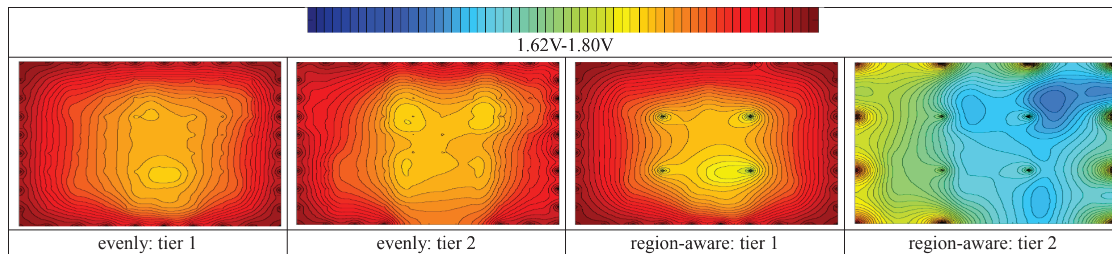

Fig. 8. The voltage maps of the benchmark S-3d-μp with different TSV planning algorithms.

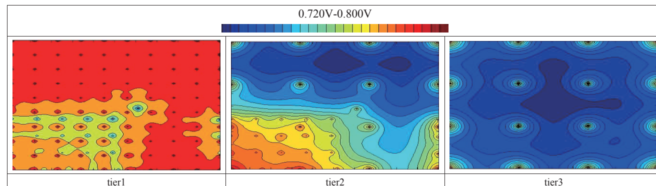

Fig. 9. The voltage maps of benchmark 1 generated with region-aware TSV planning method.

verify the validity of the generated TSV topologies. The power supply voltage is ensured while redundant TSVs are moved with the proposed region-aware planning method.

In the future, we would like to make improvements including: Taking more constraints such as thermal and timing constraints into account. Using powerful methods such as analytical methods to avoid the initialization with evenly TSV topology. Embedding a more efficient method to decide the locations where TSVs insert during one iteration, such as inserting TSVs at all concave points in the voltage map.

The main contributions of this work are summarized as follows:

1) We investigate the IR-drop problem for TSV based 3D IC more precisely with subdivision and formulate the regional distributed IR-drop constraint problem.
2) We propose the region-aware TSV planning algorithm to tackle the regional distributed IR-drop constraint problem with a non-evenly TSV topology. Furthermore, to enhance the practicability, we also take the whitespace into account when planning P/G TSVs. In comparison with the evenly TSV planning algorithm, our method can save on average 42% and 27% TSV resources without and with whitespace consideration.

References

[1] Y. Xie, G. H. Loh, B. Black, and K. Bernstein, "Design space exploration for 3d architectures," *ACM Journal on Emerging Technologies in Computer Systems*, vol. 2, no. 2, pp. 65–103, Apr. 2006.

[2] I. Savidis, S. Kose, and E. Friedman, "Power noise in tsv-based 3-d integrated circuits," *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 2, pp. 587–597, Feb 2013.

[3] X. Xiong and J. Wang, "Constraint abstraction for vectorless power grid verification," in *DAC*, ser. DAC '13. New York, NY, USA: ACM, 2013, pp. 87:1–87:6.

[4] K. Athikulwongse, A. Chakraborty, J. seok Yang, D. Pan, and S.-K. Lim, "Stress-driven 3d-ic placement with tsv keep-out zone and regularity study," in *ICCAD*, 2010, pp. 669–674.

[5] J. Yang, Z. Li, Y. Cai, and Q. Zhou, "Powerrush: A linear simulator for power grid," in *ICCAD*, nov. 2011, pp. 482 –487.

[6] Z. Feng, Z. Zeng, and P. Li, "Parallel on-chip power distribution network analysis on multi-core-multi-gpu platforms," *VLSI Systems, IEEE Transactions on*, vol. 19, no. 10, pp. 1823 –1836, oct. 2011.

[7] T. Yu, M. Wong *et al.*, "PGT-SOLVER: an efficient solver for power grid transient analysis," in *ICCAD*. IEEE, 2012, pp. 647–652.

[8] X. Xiong and J. Wang, "Parallel forward and back substitution for efficient power grid simulation," in *ICCAD*, 2012, pp. 660–663.

[9] M. Zhao, R. Panda, S. Sapatnekar, and D. Blaauw, "Hierarchical analysis of power distribution networks," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 21, no. 2, pp. 159 –168, feb 2002.

[10] G. Huang, M. Bakir, A. Naeemi, H. Chen, and J. Meindl, "Power delivery for 3d chip stacks: Physical modeling and design implication," in *Electrical Performance of Electronic Packaging, 2007 IEEE*, oct. 2007, pp. 205 –208.

[11] X. Hu, T. Toms, R. Radojcic, M. Nowak, N. Yu, and C.-K. Cheng, "Enabling power distribution network analysis flows for 3d ics," in *3DIC*, nov. 2010, pp. 1 –4.

[12] C. Zhang, V. Pavlidis, and G. De Micheli, "Voltage propagation method for 3-d power grid analysis," in *DATE*, 2012, pp. 844–847.

[13] P. Falkenstern, Y. Xie, Y.-W. Chang, and Y. Wang, "Three-dimensional integrated circuits (3d ic) floorplan and power/ground network co-synthesis," in *ASP-DAC*, 2010, pp. 169–174.

[14] H. Yu, J. Ho, and L. He, "Simultaneous power and thermal integrity driven via stapling in 3d ics," in *ICCAD*, 2006, pp. 802–808.

[15] N. Khan, S. Alam, and S. Hassoun, "Power delivery design for 3-d ics using different through-silicon via (tsv) technologies," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 4, pp. 647–658, 2011.

[16] M. Healy and S. Lim, "Distributed tsv topology for 3-d power-supply networks," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 20, no. 11, pp. 2066–2079, 2012.

[17] Z. Li, Y. Ma, Q. Zhou, Y. Cai, Y. Wang, T. Huang, and Y. Xie, "Thermal-aware power network design for ir drop reduction in 3d ics," in *ASP-DAC*, 2012, pp. 47–52.

[18] M. Jung and S.-K. Lim, "A study of ir-drop noise issues in 3d ics with through-silicon-vias," in *3DIC 2010*, 2010, pp. 1–7.

[19] W. Zhong, S. Chen, and T. Yoshimura, "Whitespace insertion for through-silicon via planning on 3-d socs," in *ISCAS 2010*, 2010, pp. 913–916.

[20] S. Tao, X. Chen, Y. Wang, Y. Ma, Y. Shi, H. Wang, and H. Yang, "HS3DPG: Hierarchical simulation for 3D P/G network," in *ASP-DAC*, 2013, pp. 509–514.

[21] Y. Chen, T. A. Davis, W. W. Hager, and S. Rajamanickam, "Algorithm 887: Cholmod, supernodal sparse cholesky factorization and update/downdate," *ACM Trans. Math. Softw.*, vol. 35, no. 3, pp. 22:1–22:14, Oct. 2008.

[22] Y. Wang, S. Yao, S. Tao, X. Chen, Y. Ma, Y. Shi, and H. Yang, "HS3DPG: Hierarchical simulation of 3D P/G Network," *submitted to IEEE Transactions on very large scale integration (VLSI) systems*.

[23] S. Yao, X. Chen, Y. Wang, Y. Ma, Y. Xie, and H. Yang, "Efficient region-aware P/G tsv planning for 3D ICs," in *ISQED*, 2014, pp. 171–178.

[24] P.-W. Luo, C. Zhang, Y.-T. Chang, L.-C. Cheng, H.-H. Lee, B.-L. Sheu, Yu-Shih-Su, D.-M. Kwai, and Y. Shi, "Benchmarking for research in power delivery networks of three dimensional integrated circuits," in *ISPD*, 2013.

[25] N. Integration and A. Modeling (NIMO) Group. Predictive technology model (ptm). [Online] http://ptm.asu.edu/.