

Partitioning-Based Approach to Fast On-Chip Decoupling Capacitor Budgeting and Minimization

Hang Li, *Student Member, IEEE*, Jeffrey Fan, *Student Member, IEEE*, Zhenyu Qi, *Student Member, IEEE*, Sheldon X.-D. Tan, *Senior Member, IEEE*, Lifeng Wu, *Member, IEEE*, Yici Cai, *Member, IEEE*, and Xianlong Hong, *Fellow, IEEE*

Abstract—This paper proposes a fast decoupling capacitance (decap) allocation and budgeting algorithm for both early stage decap estimation and later stage decap minimization in today's very large scale integration physical design. The new method is based on a sensitivity-based conjugate gradient (CG) approach. But several new techniques that significantly improve the efficiency of the optimization process were adopted. First, an efficient search step scheme to replace the time-consuming line search phase in the conventional CG method for decap budget optimization was proposed. Second, instead of optimizing an entire large circuit, the circuit is partitioned into a number of smaller subcircuits and optimized separately by exploiting the locality of adding decaps. Third, the time-domain merged adjoint method was applied to compute the sensitivity information and show that the partitioning-based merged adjoint method leads to better results than the flat merged adjoint method with the improved search scheme. Experimental results show that the proposed algorithm achieves at least ten times speed-up over similar decap allocation methods reported so far with similar budget quality, and a power grid circuit with about one million nodes can be optimized using the new method in half an hour on the latest Linux workstations.

Index Terms—Decoupling capacitor, *IR* drop, on-chip power/grid networks.

I. INTRODUCTION

POWER integrity has become the most insidious issue in today's deep submicrometer and nanometer very large scale integration (VLSI) regime. *IR* drop is caused by device switching current flowing through the parasitic power/ground network. A supply voltage (VDD/GND) variation throughout the entire chip will occur, which will lead to an adverse impact on chip performance, longer path delay, or even logic

failure. With reduced noise margin and increased switching frequency as technology scales, it is required to confine the supply voltage fluctuation within a certain range of nominal VDD value to guarantee reliable power delivery. For removing dynamic *IR* drop that arises from resistive and inductive effects, intentionally adding decoupling capacitance (decap) between power and ground buses or between power/ground buses and substrate is the most efficient way. As shown in Fig. 1, decaps provide a reservoir of current that is instantly available for the near switching components to remove spikes and glitches in the power rail. Since on-chip decaps are typically manufactured using MOS transistors, and excessive on-chip decaps could cause more leakage power, low yield, and lower resonant frequency [1], the total decap area should be added in an area-efficient way.

Budgeting decap in an area-efficient way, however, is a difficult task due to prohibitive analysis costs of power/grid (P/G) networks with millions of nodes and extracted on-chip and off-chip *RLC* components in modern VLSI design. Mathematically, optimal decap allocation is a nonlinear optimization problem, and many existing approaches [7], [18] use sensitivity-based optimization methods to solve the problem. To compute the sensitivity, transient simulations of the whole P/G networks have to be carried out at every optimization step. Given the fact that the transient simulation of P/G networks with millions of nodes is already an extremely time-consuming task, the CPU and memory cost of the optimization method that uses transient simulations as internal loops will be prohibitive. Recent study [7] shows that allocating decaps for a P/G grid with about one million nodes will take about 10 h in modern workstations with improved simulation techniques. Given the increasing sizes of P/G networks, existing decap budgeting techniques do not scale well for future VLSI on-chip power distribution network design and verification.

An alternative approach based on the multigrid concept was proposed by Wang and Marek-Sadowska in [21]. But this method is limited to the regular-mesh or regular-mesh-like P/G structures. This method exploits the geometrical multigrid method for sizing the widths of P/G wires and allocating decaps. At the reduced mesh level, sequential quadratic programming is still used, where sensitivity is computed by the adjoint method only. But for nonmesh structured P/G networks, this method may not work due to the use of a geometrical multigrid idea. Also, during the re-mapping process (map the decap results from high level grid to low level grids), the method requires that all the widths of P/G wires be same in a

Manuscript received March 30, 2005; revised June 29, 2005 and September 8, 2005. This work was supported in part by the National Science Foundation under CAREER Award CCF-0448534 and Grant OISE-0451688 and by UC MICRO #04-088 via Cadence Design System, Inc. The work of Y. Cai and X. Hong was supported by the Hi-Tech Research and Development (863) Program of China under Grant 2004AA1Z1460 and the National Natural Science Foundation of China (NSFC) under Grant 60476014. This paper was presented in part at the *Proceedings of the Design Automation Conference (DAC'05)* and the *Proceedings of the International Symposium on Quality Electronic Design (ISQED'05)* [11], [14]. This paper was recommended by Associate Editor S. Sapatnekar.

H. Li, J. Fan, Z. Qi, and S. X.-D. Tan are with the Department of Electrical Engineering, University of California, Riverside, CA 92521 USA (e-mail: hli@ee.ucr.edu; jfan@ee.ucr.edu; stan@ee.ucr.edu).

L. Wu is with Cadence Design Systems Inc., San Jose, CA 95134 USA.

Y. Cai and X. Hong are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: caiyc@mail.tsinghua.edu.cn; hxl-dcs@mail.tsinghua.edu.cn).

Digital Object Identifier 10.1109/TCAD.2006.870862

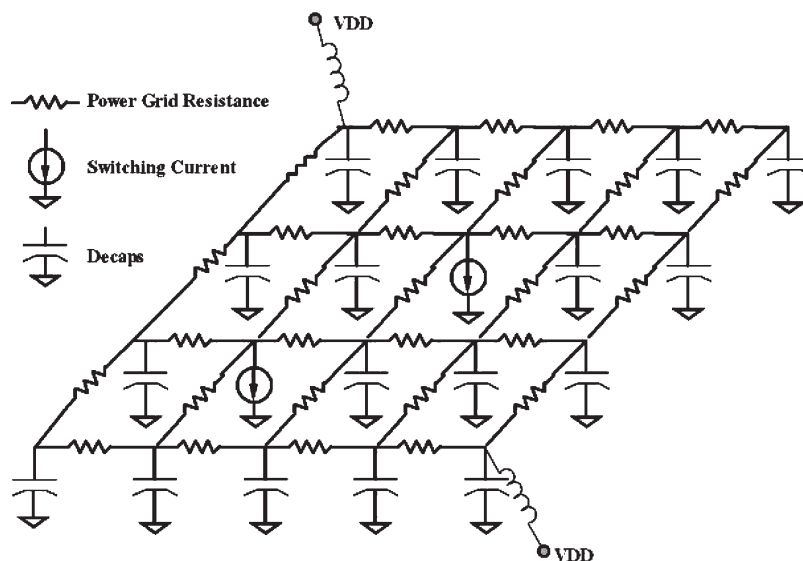


Fig. 1. Model of power grid networks.

P/G strip. Otherwise, the remapping process will not guarantee the existence of the solution at the low level. Actually, this will not be satisfied for actual industry P/G network circuits. Also, if there are many C4 pads (voltage sources), the reduction ratio will be reduced, so does the efficiency of this method. In short, the multigrid-based method has many restrictions and limitations.

In this paper, we propose a general fast decap allocation and budgeting algorithm in solving a large power grid circuit, which is modeled as *RLC* networks considering both on-chip and off-chip parasitics. The new method is based on sensitivity-based conjugate gradient (CG) methods. Our contributions include the following: 1) Instead of doing line search at every step in CG optimization, we develop a new search step scheme to speed up the optimization process; 2) Based on the local effect of adding decap to reduce *IR* drops, we partition the whole circuit into a number of subcircuits and optimize them individually. A noise-aware partition scheme is proposed to perform the required partitioning; and 3) We apply the time-domain merged adjoint network method in [7] and [18] for sensitivity calculation in the partitioning-based optimization framework. We show that combining the proposed partitioning scheme with the merged adjoint method leads to a much faster optimization process with similar solution quality given by the flat CG using the merged adjoint method [7], [18]. The combination of our new optimization algorithm and partition scheme significantly improves the analysis speed for very large power/ground circuits even on a single CPU workstation.

We notice that adding decaps alone may not solve all the voltage drop problems in power/ground grids. Our recent study considering leakage of decaps [8] shows that decap leakage can lead to more decap areas for the same noise reduction effects. In this case, considering both wire sizing and decap allocations are better solutions [8]. The proposed method can be used to optimize the power/ground networks before or after the wire sizing step, as leakage currents are mainly dc currents and can be optimized more efficiently by wire sizing.

We notice that the time-domain merged adjoint method has been used for transistor sizing [6] and then later applied to the decap optimization in [7] and [18]. But, we show in this paper that with the proposed partitioning scheme the merged adjoint method works better than the simple application of the merged adjoint method in the CG optimization framework.

The rest of this paper is organized as follows. Section II briefly reviews existing sensitivity-based decap budgeting algorithms. In Section III, we describe our new algorithm, with theoretical analysis regarding time efficiency. Section IV gives a description of a graph-based partition algorithm dealing with a special decap situation. The experimental results are summarized in Section V to verify our method, with conclusions in Section VI.

II. REVIEW OF PREVIOUS DECAP OPTIMIZATION ALGORITHMS

Existing on-chip decap budgeting algorithms basically fall into two categories. In [2], [12], [16], and [17], the current pattern around hot spots (where violation of *IR* drop occurs) is derived and the amount of electric charge needed to supply that current demand is estimated. To obtain an optimal decap budget, a critical step for these methods is the precise estimation of voltage drops, which unfortunately proves to be difficult for practical P/G networks without simulation.

Another category is sensitivity-based approaches based on actual circuit simulation, which is a more accurate and well-accepted method. Fig. 2 gives an illustration of VDD fluctuation of a node within one clock cycle [18]. The violation area at node j is defined as

$$g_j(c_1, \dots, c_n) = \int_0^T \max(V_{\min} - v_j(t), 0) dt \quad (1)$$

which equals the shaded area below a certain VDD threshold in the graph. The sensitivity of decap added at node i in

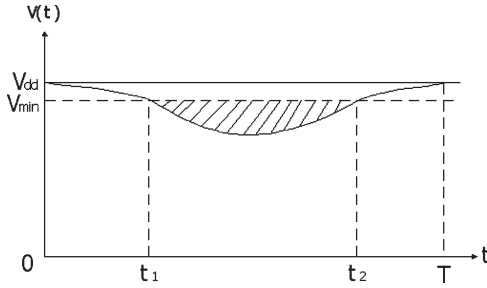


Fig. 2. Illustration of IR drop violations.

contributing to remove this violation area at node j is

$$s_{ij} = \frac{\partial g_j(c_1, \dots, c_n)}{\partial c_i} \quad (2)$$

where c_i is the decap value added at node i and n is the number of nodes where decaps can be added. If we allow all circuit nodes to be candidates for adding decaps, n equals the number of circuit nodes. The adjoint network method is the conventional way to calculate this sensitivity, which requires two full circuit simulations, one for the original network and the other for the adjoint network. The sensitivity for capacitive elements can then be expressed as

$$s_{ij} = \int_0^T v'_{i,j}(T-t) \dot{v}_i(t) dt, \quad i = 1, 2, \dots, n \quad (3)$$

where $\dot{v}_i(t)$ is the derivative of the voltage waveform at node i in the original network and $v'_{i,j}(T-t)$ is the waveform at node i in the adjoint network under unit step current excitation at violation node j .

In previous work [18], a sensitivity-based algorithm is developed to optimize the allocation of decaps in a standard cell design format. In its problem formulation, it minimizes the objective function

$$\sum_{j=1}^m g_j(c_1, \dots, c_n) \quad (4)$$

which is the sum of the violation areas of all the violation nodes. The constraints are the lower and upper bounds of the decap values and the total decap areas in each row (the total white space in the layout) for each decap node. n is the number of violation nodes. Then the problem is solved by a quadratic programming solver. The iteration continues until all violations are eliminated.

For gradient computation, the adjoint network method in (3) is used to obtain the sensitivity value of the objective function to each decap value. The sensitivity of every violation node with respect to every decap node is computed within each iteration. Therefore, the time complexity is approximately $O(n^{1.5}lmh)$, where m is the number of violation nodes, l the number of sampling time points, and h the number of optimization iterations. The term h depends on the convergence rate of the optimization method and $n^{1.5}$ is the typical time complexity for

solving sparse $n \times n$ matrices [19]. It is clear that the speed of this algorithm depends on the number of violation nodes and will become intolerable for extremely large circuits.

Recent work [7] improved the sensitivity calculation by applying the time domain merged adjoint network method, which will be covered later in the next section. The decap area is incorporated into the objective function as

$$\left(\sum_{i \in N_{\text{decap}}} c_i \right) + \alpha \left(\sum_{j \in N_{\text{vio}}} g_j \right) \quad (5)$$

where N_{decap} is the set of all the decap nodes and N_{vio} is the set of all the violation nodes. The new objective function is solved by the traditional CG method within each optimization iteration. The weighting factor α in (5) is used to balance the two terms and will keep changing in each CG iteration. However, such a balance could be misleading since the optimization direction may be decided by the value of α rather than the sensitivity of each decap value to the objective function. An improper α will cause extra line searches, or even optimization failure. What is more, each line search during CG optimization also requires the evaluation of the objective function at the cost of a full transient simulation, which renders the algorithm inefficient and unreliable.

In this paper, we present our improved CG (iCG) algorithm, which takes the advantage of the fast sensitivity computation in [7], while avoids the ambiguity of the old optimization objective and the inefficient line search phase.

III. ICG ALGORITHM

In our new iCG algorithm, we follow the same problem definition in [18]. Specifically, we formulate the optimization problem as

$$\text{objective function : } \min \sum_{j=1}^m g_j(c_1, \dots, c_n) \quad (6)$$

$$\text{subject to constraints : } c_i \leq d_i, \quad d_i \geq 0 \quad (7)$$

where m is the number of violation nodes and n the number of decap nodes. d_i is the maximum decap allowed at node i , a parameter decided by the available layout white space around node i . Notice that the upper bounds d_i at different nodes may not be independent as shown in the standard cell-based layout [18] since total decap areas may be fixed. But those dependences can be easily considered in our algorithm as those geometry-related constraints are explicitly checked and enforced during the optimization process.

Although we do not put decap areas in the objective function as in [7] because they cause many problems in the CG-based optimization framework as discussed in previous section, we do minimize the decap area implicitly in the CG optimization framework by following the CG direction in each CG iteration step and later performing the proposed binary line research (to be discussed later).

Note that the decap budgeting problem defined in (6) and (7) can be used in the early floor-planning stage for decap budget estimation, where d_i is the available white space that can be used for adding decaps. The same optimization formulation can also be used in the later stage of physical design to minimize the existing decap budget, since some decaps are placed heuristically in the previous design stage. d_i then represents the decap value already allocated at node i .

As mentioned in Section II, the sensitivity of the objective function (6) with respect to each decap value can be computed by the merged adjoint network method. In a conventional adjoint network, a unit step current source is placed at each violation node when calculating the sensitivity described in (2). However, since all the adjoint networks for each violation node share the same topology, we may combine the step current sources together based on circuit superposition. And the sum of all the decap sensitivities would turn into

$$\sum_{j=1}^m s_{ij} = \int_0^T (v'_{i,\text{all}}(T-t)) \dot{v}_i(t) dt, \quad i = 1, 2, \dots, n \quad (8)$$

where $v'_{i,\text{all}}(T-t)$ is calculated from the merged adjoint network with combined step current sources. The merit of this method can be observed immediately since

$$\sum_{j=1}^m s_{ij} = \frac{\partial \sum_{j=1}^m g_j(c_1, \dots, c_n)}{\partial c_i} \quad (9)$$

where the right-hand-side is just the sensitivity of the objective function we need in our problem formulation. Therefore, all the decap's sensitivity to the objective function can be calculated in only two transient simulations, which greatly improves the efficiency of the algorithm.

Although the merged adjoint method can significantly reduce the number of full circuit transient simulations, a number of simulations are still needed at each step in the CG method. The reason is that the direct application of the CG method in [7] requires evaluations of the objective function at different points along the current gradient direction to find the minimum cost in the objective function. But the cost of simulation in this line search phase is expensive and could offset the efficiency gained by the merged adjoint method.

In order to avoid this, we develop a simple yet efficient search step computation method. The method is based on the observation that the step size in each search direction can be simply determined by computing the maximum decap value allowed on one or some nodes under this search direction. In other words, we determine the maximum step we can take in the current direction and set each decap value according to this step. We will continue to do so until a violation criterion has been met. One problem with such a maximum-allowable-step scheme is that it may overestimate the decap areas. To alleviate this problem, a binary search will be performed to find the best step such that all violations are just removed and decap areas are minimized. The main difference of our method from the traditional CG method is that we only need to do a binary search just once in the entire optimization process, which is in contrast

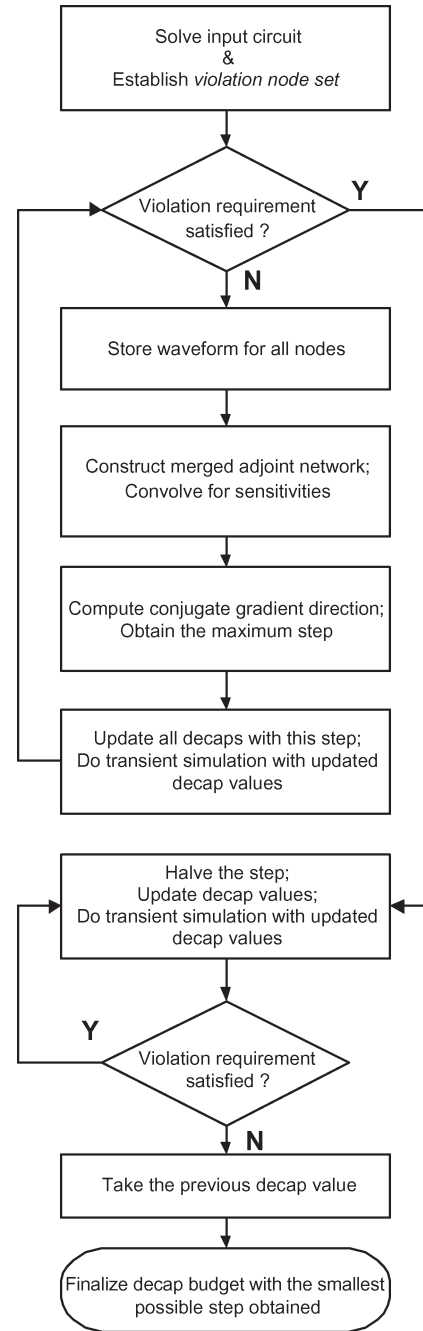


Fig. 3. Proposed decap budgeting optimization flow.

to the traditional CG method, where the line search is carried out at every CG step during optimization.

The new proposed decap budgeting algorithm can be illustrated in Fig. 3, which has two stages. First, the existence of a solution is checked in the present direction computed by the CG method. Then, a solution is located by binary searches. Within each CG iteration, only two transient simulations are needed for gradient computation. So the time complexity of the algorithm is about $O(n^{1.5}l(h+r))$. Again, n is the circuit node count, l is the number of time points, and h is optimization iteration number. r denotes the number of steps in which a binary search is attempted.

The time complexity in [7] is derived in the original paper and can be expressed as $O(n^{1.5}lh(2+r'))$, where all variables are defined as above except r' , which is the number of line searches in each iteration. Usually, h , r' , and r are all within the same order of magnitude. So the proposed algorithm can be expected to be much faster as we only do the binary (line) search once and avoid the hr' term in the time complexity. Compared with the decap algorithm in [18], the proposed method also leads to a great run time reduction, as it is independent with the number of violation nodes. Also, the faster convergence and less computation cost of CG than quadratic programming contributes to the time efficiency as well.

IV. PARTITIONING-BASED ANALYSIS OF P/G NETWORK AND DECAP BUDGETING

Accurate decap budgeting relies heavily on the repeated transient analysis of the entire power grid. With a full-chip power grid model consisting of millions of nodes and elements, the burden on computation and memory storage during optimization is huge. For decap optimization, which asks for full circuit simulation at each time point, it is desirable to reduce the problem size of analysis by partitioning the large circuit into several smaller subcircuits, and optimize decap budget of each partition separately, or even in parallel.

The partitioning-based strategy is also supported by the fact that adding decap to remove the IR drop violation is a local effect. Normally, the violation occurs in the center or a certain area on the chip. For the emerging flip-chip packaging technology, this local effect becomes more prominent [4]. The number of violation nodes compared to the total circuit node number is also supposed to be small. Otherwise, the entire power network should be redesigned. For each violation node, the most effective decap locations are always the nodes close to it. Thus, we should expect only a few good decap candidate locations through the entire chip for each violation node.

A. General Partition Algorithm

In our paper, we use the graph-based multilevel minimum cut algorithm for the partitioning task [9], which provides an extremely fast speed on large graph sizes, where a graph with one million vertices can be processed roughly within 1 min. This ensures that the partition phase will not bottleneck our entire partitioning-based optimization flow. Specifically, we first convert the power grid netlist in simulation program with integrated circuit emphasis (SPICE) format to the graph file style accepted by METIS. Then, we read the output file generated by METIS and create each individual partition's netlist file. These files are processed serially by our iCG optimization engine to have the individual partition decap budget result. The final decap budget will simply be the combination of all partition's budgets.

A direct use of partitioning without considering the IR drop violation could have adverse impacts on the entire decap optimization. We should avoid putting violation nodes as the boundary nodes, whose node voltage cannot be reduced by adding decaps, since boundary node voltages are treated as independent voltage sources as discussed below.

TABLE I
DECAP BUDGET COMPARISON BEFORE/AFTER PARTITION

Subcircuit Name	Original Budget	w/ boundary		w/o boundary	
		budget	deviation	budget	deviation
1	1.00	1.11	11.2%	1.48	48.1%
2	1.00	0.77	-22.9%	1.64	64.0%
17	1.00	1.22	22.6%	1.23	22.6%
18	1.00	1.38	38.8%	1.72	71.7%

B. Boundary Condition for Subcircuits

For each subcircuit, we need to consider the influence of other subcircuits on its decap budget optimization. We call this boundary condition of the subcircuit. If the boundary condition for each individual partition is not considered, the decap budget will be conservative compared to the one presented in the original netlist. The reason is that the subcircuit, as opposed to the full circuit, confines currents to flow in a smaller area instead of the global P/G network. As consequence, more currents (due to smaller resistance to true ground) occur in the same area after partitioning. These overestimated currents will cause greater IR drops than actual ones.

To solve this problem, we keep the boundary node waveforms in piecewise linear (PWL) form derived from a full circuit transient simulation at the beginning of the optimization. In this way, the voltage waveforms are the same even when each subcircuit is simulated independently.

In Table I,¹ a circuit with 240 K nodes is partitioned into 20 pieces. Decap budgets are compared between several subcircuits with PWL boundary conditions and the ones without. Deviation is calculated as the difference between the budget of each subcircuit and the budget of the same subcircuit in the original circuit optimization. We notice that there are only four subcircuits containing violation nodes, and the deviation of the simple partition budget without the boundary conditions is much greater than the one with boundary conditions.

C. Noise-Aware Partitioning (NAP)

Keeping the boundary node voltage as independent voltage sources introduces another problem. The violation nodes may appear as the boundary node in each partition, and there is no way to reduce the voltage drop of the PWL voltage source at the violation node by adding decaps. Therefore, we should explicitly avoid putting violation nodes into the boundary node set during partitioning. This can be easily implemented by assigning a relatively heavy edge weight incident on each violation node as well as the nodes adjacent to them. Since the partition algorithm attempts to minimize the total cut weights on the boundary, those weighted violation node edges have a very low possibility to be cut.

Another issue is that we should keep the good decap candidate nodes for violation nodes in the same subcircuit. Otherwise, it will be more expensive to reduce the IR drops of the violation nodes by using less effective nodes available in a

¹In this table, as well as in other tables in the rest of this paper, the decap budgets for the comparison of CG, iCG, and later partitioning-based iCG algorithms are scaled values such that the decap values given by CG using the merged adjoint method is 1.0.

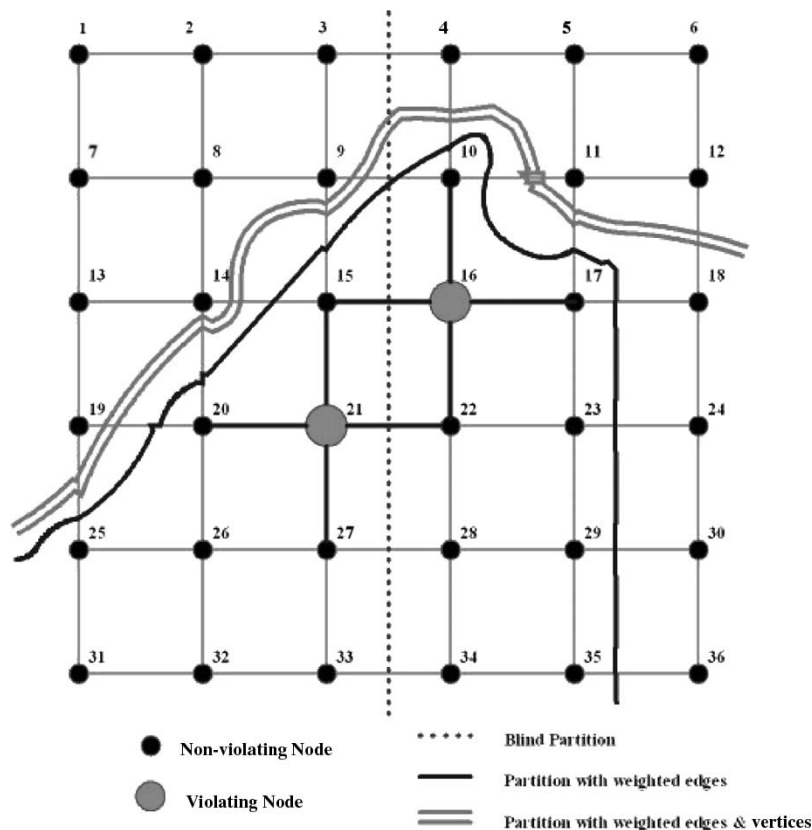


Fig. 4. Example of noise-aware partition scheme.

subcircuit. To this end, we need to consider adding decap range for violation nodes during partitioning. The current sources at the violation nodes are typically the cause of IR drop violation and will be assigned into one subcircuit along with all the nodes they are directly connected to or in nearby locations. To achieve this, we can assign a relatively small vertex weight to each violation node containing the current source, as well as a predefined radius, within which the nearest nonviolation nodes close to it are also given the same weight. Given the fact that the partitions should be balanced in terms of each partition’s total vertex weight, the violation nodes will be aggregated with a host of nearby nonviolation decap candidate nodes.

Fig. 4 gives a simple example of the above-proposed partition scheme. We observe that the violation nodes 16 and 21 are easily separated into two partitions without any differentiation between them and the nonviolating ones. After we add larger weights on the edges around them, they are captured into one partition. If we further assign a smaller vertex weight to them, as well as the ones adjacent to them, which are nodes 10, 15, 17, 20 and 22, more surrounding nodes will go with nodes 16 and 21 into the same subcircuit, which is exactly what we expected.

By using this NAP scheme, we did the test on the previous 240 K circuit again. The original budget refers to the budgets obtained in the flat run of the optimization for each subcircuit.

We observe that the results in Table II show that the new partition budget is very close to or even smaller than the one from the original circuit. In Table II, original budget is obtained by using CG with the merged adjoint method. The reason is related to the merged adjoint method, which leads to worse

TABLE II
DECAP BUDGET COMPARISON UNDER NAP

Subcircuit Name	Original Budget	Partitioned Budget
4	1.0	0.80
5	1.0	0.63
14	1.0	1.01

results compared to the CG method using the sensitivities of individual violation node area (without using the merged adjoint method). The merged adjoint method computes the sensitivity of the objective function, which is the sum of all the violation node areas, with respect to each specific decap instead of sensitivities with respect to each individual violation node area. As a result, the CG with the merged adjoint method will lead to worse results than the CG using the sensitivities of individual violation node simply because the later method can use each decap to its maximum effects for reducing voltage drops while the former method does not have the flexibility to fine tune each decap’s contribution to a specific violation node. Our recent study by using a sequence of linear programming (SLP) methods for decap budgeting shows that using individual sensitivities of violation decap nodes can get a much better quality than the merged adjoint method [13].

On the other hand, the partitioning scheme tends to reduce the adverse effects of the merged adjoint method as partitioning tries to make the sensitivities to be limited to the violation nodes in each partition. In the extreme case, each partition has just one node, we go back to the nonmerged sensitivity computation case in which we should have the best result in this regard. But

in this case, another problem will kick in: a very large partition number will lead to no solution for our iCG optimizer due to the limited decap nodes in each partition. Therefore, for a certain range, the decap budget from a larger partition number will result in better optimization results.

Although partition helps to improve the optimization results in a CG algorithm using the merged adjoint method, it also degrades the optimization quality as it makes some violation nodes harder to optimize, which will be explained in Section V. As a result, the net effect of using the partitioning scheme is the increased efficiency of the optimization with similar quality of the CG method using the merged adjoint method [7].

Note that for the power/ground networks with many C4 pads, the C4 region can be used as natural partitions as shown in [4]. In this case, no explicit partitioning is required and the partitioning-based decap optimization can be done naturally in C4 power/ground networks.

D. Partitioning-Based Decap Optimization Flow

The whole partitioning-based optimization flow is given in Fig. 5. It performs only two full circuit transient simulations, one at the very beginning to report all the violation nodes and record the original boundary node waveform, and another in the last to verify the optimization result. Compared to many full simulations carried out in a flat run mode, the time overhead from these two full circuit simulations becomes less significant. The rest of the simulations will be conducted on each subcircuit only, which are much faster than full circuit simulation even done sequentially.

We assume a number of N partitions (subcircuits) with approximately the same amount of nodes in each partition. Since the partitions are processed sequentially in our experiment, and parameters l , h , and r will not change too much with partitioning, the new time complexity becomes

$$N \left[\left(\frac{n}{N} \right)^{1.5} l(h+r) \right] = \frac{[n^{1.5}l(h+r)]}{\sqrt{N}}. \quad (10)$$

Therefore, the time complexity of the partitioning-based optimization algorithm will decrease with the square root of the number of partitions. However, the number should not be very large either because a very small subcircuit could result in no solution of optimization due to reduced decap candidate positions. In case no solution is found in the subcircuits, we will halve the partition number and rerun the whole algorithm again for a relaxed partition area. Since we snapshot the IR drop violation by defining an effective area for adding decap in partitioning, the optimization effect is supposed to be guaranteed when all the partitions are combined together.

If parallel computing is allowed, the time complexity can be further reduced to

$$\frac{[n^{1.5}l(h+r)]}{N^{1.5}} \quad (11)$$

since there is no communication needed between different partitions during decap optimization.

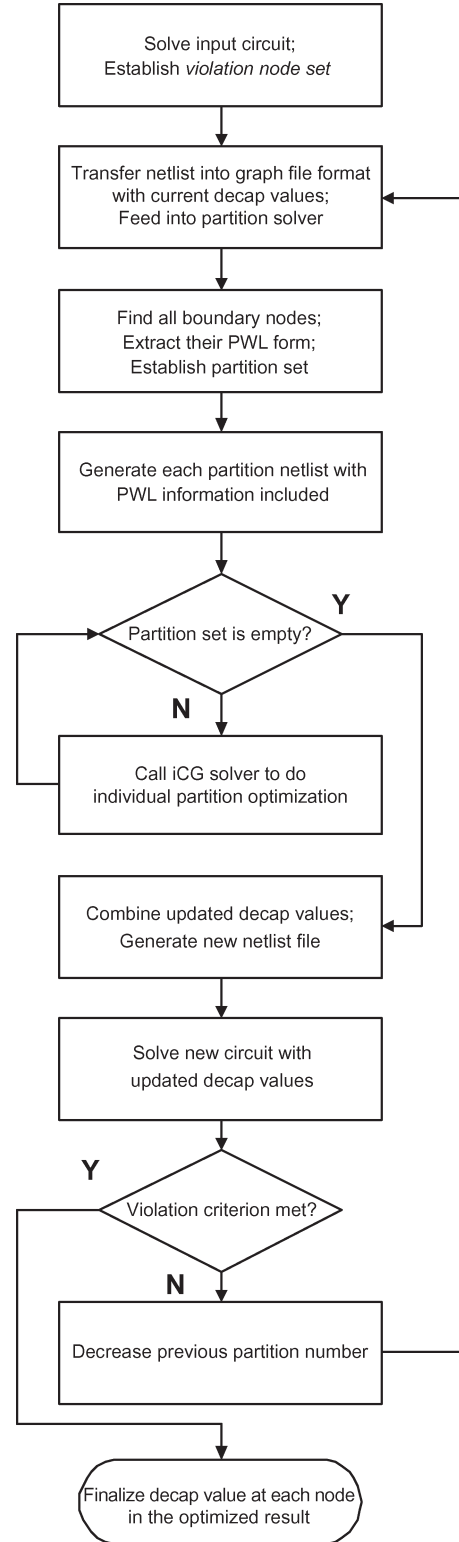


Fig. 5. Partitioning-based optimization flow.

V. EXPERIMENTAL RESULTS

We implement our proposed algorithm in C++. All experiments are carried out on a Linux PC with dual 3.0-GHz Xeon CPUs and 2-GB memory. All test circuits are generated by the authors with realistic parameters for R , C , and current sources

TABLE III
COMPARISON WITH EXISTING CG METHOD FOR P/G DECAP OPTIMIZATION

Circuit	#Nodes	#Vio Nodes	CG1 (existing)			CG2 (proposed)			speedup ratio
			decap	iter	time(s)	decap	iter	time(s)	
ckt1	88	29	1.00	8	2.3	1.16	1	0.1	23
ckt2	336	63	1.00	10	15.2	1.32	2	0.8	19
ckt3	1,233	143	1.00	10	132	1.08	1	2.4	55
ckt4	12,673	1,083	1.00	8	1,995	1.18	1	54	37
ckt5	89,496	592	1.00	5	7,241	1.59	1	394	18

TABLE IV
COMPARISON BETWEEN FLAT AND PARTITIONING-BASED DECAP OPTIMIZATION

Circuit	# of nodes	# of vio nodes	CG1		CG2		Partitioned CG2			
			budget	time(s)	budget	time(s)	partition no.	budget	time(s)	speedup ratio
ckt6	242,600	49,626	1.00	9,592	1.29	1,746	5	1.55	744	13
	-	-	-	-	-	-	10	1.14	713	13
	-	-	-	-	-	-	20	1.03	438	22
ckt7	421,320	26,843	1.00	15,555	1.14	1,370	10	2.09	1,077	14
	-	-	-	-	-	-	20	1.87	1,034	15
	-	-	-	-	-	-	40	1.09	765	20
ckt8	827,025	87,903	N/A	N/A	N/A	N/A	20	1.00	2,619	N/A
	-	-	-	-	-	-	40	0.61	1,711	N/A
	-	-	-	-	-	-	80	0.60	1,705	N/A
ckt9	1,004,960	67,105	N/A	N/A	N/A	N/A	25	1.00	2,812	N/A
	-	-	-	-	-	-	50	0.54	2,675	N/A
	-	-	-	-	-	-	100	0.49	2,093	N/A

based on industry designs. The off-chip inductive parasitic effects are also considered. Some figures are exaggerated in order to test the versatility of our algorithm. For each test case, we artificially set the power noise level such that the number of violation nodes presented in the circuit is within the 20% range of the total node count. Keep in mind that we cannot count solely on adding decaps to eliminate all IR drop violations and a huge amount of violation is not reasonable for decap solution.

We first compare our method with that in [7] for small circuits without partitioning. To make comparison possible, we implement that in [7] in such a way that before each line search an explicit attempt to bracket the minimum is made, and if the minimum is found to lie at the start of the line, α is augmented. In this way, we avoid the problem mentioned in Section II and make the algorithm robust enough for all our tests.

Table III summarizes the comparison, where CG1 denotes the method in [7] and CG2 denotes our iCG method. Columns 1, 2, and 3 represent circuit names, total node numbers, and violation node numbers, respectively. Parameters including voltage drop tolerance and the maximum decap at each node can be specified by users and are the same for both methods. The last column compares the total optimization CPU time for the two algorithms. For all these circuits, violation elimination requirement is successfully achieved after both decap optimizations. The CPU time efficiency of the proposed method is usually more than ten times faster than the method in [7]. We also notice that the new method does trade some qualities for speed-up.

We apply our partitioning-based optimization algorithm for larger power grid circuits. As shown in Table IV, comparisons are made between the budget and CPU time of the flat CG1, CG2, and partitioning-based CG2 algorithm. Please note that the CPU time includes initial simulation time and partition time. The circuit sizes range from 240 K to 1 M nodes. While CG1

is still capable of solving ckt6 and ckt7, it fails to work on 800 K and 1 M cases. The main reason arises from the memory limitation (we have 2G memory in our Linux workstation) for LU decomposition and waveform storage in sensitivity calculation during each transient simulation. CG2 suffers the same problem when doing the optimization flatly. The partitioning-based algorithm, on the other hand, can handle these cases very easily.

As can be seen from Table IV, the partitioning-based algorithm optimizes all circuits successfully, and the budget achieved is comparable, or even smaller than the flat optimization runs. The time advantage is also impressive. The circuit with one million nodes can be optimized in about half an hour, as opposed to a 10-h run time in [7] for the same circuit volume.² Another thing that needs to be pointed out is that we simulate the circuits based on direct LU decomposition, while a structure level reduction technique is applied in [7] for the simulation on an actually smaller circuit size. The time efficiency of the algorithm is therefore more obvious.

Since the efficiency of our proposed method depends on the linear solver used, any speedup techniques like hierarchical approach [22], multigrid methods [10], iterative methods [3], model reduction methods [5], [20], and random walk-based algorithm [15] can be used to speed up and increase the capacity of the proposed method.

We also notice the difference among various partition sizes. For each example, the larger the partition number, the faster the speed as we projected. However, the decap budget experiences a variation. As an example of the 400 K circuit illustrated in Fig. 6, with small partition numbers, the budget is over-estimated compared to the flat-run result. The reason is that

²The CPUs used in two papers are different (Intel Xeon versus Sun UltraSparc). It is difficult to scale the CPU times in terms of raw clock speeds.

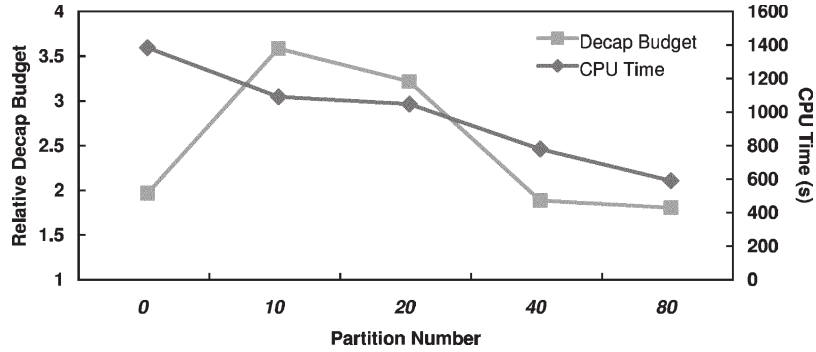


Fig. 6. Comparison of decap budget and CPU time between different partition sizes.

TABLE V
COMPARISON BETWEEN SLP AND PARTITIONING-BASED OPTIMIZATION

Circuit	# of nodes	# of vio nodes	SLP (sen = 0.1)		SLP (sen = 1)		Partitioned CG2				
			budget	time(s)	budget	time(s)	partition no.	METIS time(s)	budget	time(s)	speedup ratio
ckt10	6,105	206	1.00	470	0.15	549	1	N/A	0.20	41	11
	-	-	-	-	-	-	3	< 1.0	0.27	25	19
	-	-	-	-	-	-	5	< 1.0	0.37	20	24
ckt11	29,425	181	1.00	1,871	0.10	2,123	1	N/A	0.90	140	13
	-	-	-	-	-	-	30	< 1.0	0.28	62	30
	-	-	-	-	-	-	50	< 1.0	0.13	59	32
ckt12	89,496	251	1.00	6,911	0.10	6,807	1	N/A	0.87	929	7
	-	-	-	-	-	-	100	< 1.0	1.23	185	37
	-	-	-	-	-	-	200	< 1.0	1.63	176	39
ckt13	123,280	301	1.00	10,700	0.10	10,557	1	N/A	0.75	1,124	10
	-	-	-	-	-	-	100	< 1.0	1.30	259	41
	-	-	-	-	-	-	200	< 1.0	1.11	249	43
	-	-	-	-	-	-	300	< 1.0	1.15	240	45
ckt14	536,705	373	N/A	N/A	N/A	N/A	1	N/A	1.00	4,052	N/A
	-	-	-	-	-	-	100	2.0	0.47	1,226	N/A
	-	-	-	-	-	-	300	2.0	0.47	1,102	N/A
	-	-	-	-	-	-	400	2.0	0.52	1,093	N/A

partitioning makes some violation node area harder to remove, as the available candidate decap nodes become less, due to the formation of subcircuits and PWL voltage sources as the boundary. Those nodes are typically close to the boundaries of subcircuits. Therefore, a larger decap value is added in each optimization iteration.

But as the partition number increases, the decap budget will drop as shown in Table IV. The reason has actually been explained in Section IV-C. The results in Table IV further give evidence that the partitioning scheme can alleviate the adverse effect brought by the merged adjoint method. With larger partition numbers, we are more close to the nonmerged sensitivity computation case, so we will get better result.

Since partitioning itself also degrades the decap allocation quality as discussed above, the total decap allocation quality, due to two effects, leads to similar quality given by the CG using the merged adjoint method, as shown in Table IV, but the proposed method can be much faster and more capable than the flat CG method.

Also, the partition number should not be too large. Otherwise, the number of decap nodes will become too small to have a solution for removing violating areas in the subcircuit. In such cases, we reduce the number of partitions and more partition iteration will be conducted as mentioned in Fig. 5, leading to a longer run time.

Furthermore, we make a comparison between the partitioning-based decap optimization algorithm and SLP [13], as shown in Table V. In the SLP method, the sensitivities of each violation node voltage with respect to all the decaps are computed (as a result, it becomes more expensive). So SLP enjoys the most flexibility during the decap optimization and gives better results.

But the results of SLP strongly depend on a tuning parameter called sensitivity parameter [13]. Given a fixed sensitivity parameter (i.e., sen = 0.1), the experiment shows roughly ten times or more speed-up of the proposed method over the SLP method at the cost of decap quality in some circuits. Please note that the SLP method fails to solve the circuit with 500 K nodes, while the partitioning-based algorithm is able to solve the circuit in a timely manner. The increase of the sensitivity parameter will improve the quality of decap budget at the expense of slower speed in some circuits. Hence, the partitioning-based algorithm is more applicable when speed is the most concern for designers. One may notice that the METIS partition time is trivial in comparison to the overall CPU time.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a fast decap optimization solution, targeting circuits with very large size. The combination

of our proposed iCG algorithm and the partitioning-based optimization scheme can efficiently optimize power grid circuits with million nodes in a timely manner. Our theoretical analysis on the time complexity shows that the new algorithm outperforms most of the existing decap allocation algorithms. Practically, we show that combining the partitioning scheme with the merged adjoint method leads to a faster optimization process without loss of the optimization quality compared to the flat CG algorithm with the merged adjoint method. Experimental results on a number of power grid circuits demonstrate that the proposed algorithm achieves roughly at least ten times speed-up or more over similar decap allocation methods reported so far, and the power grid circuit with about one million nodes can be optimized in about half an hour on the latest Linux workstation.

In the future, more efficient circuit simulation techniques like hierarchical approach [22] and model reduction approaches [20] will be used to improve the transient simulation of power/ground networks. Also, a parallel simulation will be explored to further improve the efficiency of the decap budgeting algorithm.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers of this paper for their comments and suggestions that helped improve this work.

REFERENCES

- [1] S. Bobba, T. Thorp, K. Aingaran, and D. Liu, "IC power distribution challenges," in *Proc. ICCAD*, San Jose, CA, 2001, pp. 643–650.
- [2] H. H. Chen and D. D. Ling, "Power supply noise analysis methodology for deep-submicron VLSI chip design," in *Proc. DAC*, Anaheim, CA, 1997, pp. 638–643.
- [3] T. Chen and C. C. Chen, "Efficient large-scale power grid analysis based on preconditioned Krylov-subspace iterative method," in *Proc. DAC*, Las Vegas, NV, 2001, pp. 559–562.
- [4] E. Chiprout, "Fast flip-chip power grid analysis via locality and grid shells," in *Proc. ICCAD*, San Jose, CA, Nov. 2004, pp. 485–488.
- [5] E. Chiprout and T. Nguyen, "Power analysis of large interconnect grids with multiple sources using model reduction," in *Proc. Eur. Conf. Circuit Theory Design*, Stresa, Italy, Sep. 1999, pp. 433–436.
- [6] A. R. Conn, R. A. Haring, and C. Visweswariah, "Noise considerations in circuit optimization," in *Proc. ICCAD*, San Jose, CA, 1998, pp. 220–227.
- [7] J. Fu, Z. Luo, X. Hong, Y. Cai, S. X.-D. Tan, and Z. Pan, "A fast decoupling capacitor budgeting algorithm for robust on-chip power delivery," in *Proc. ASPDAC*, Yokohama, Japan, Jan. 2004, pp. 505–510.
- [8] —, "VLSI on-chip power/ground network optimization considering decap leakage currents," in *Proc. ASPDAC*, Shanghai, China, Jan. 2005, pp. 735–738.
- [9] G. Karypis, R. Aggarwal, and V. K. S. Shekhar, "Multilevel hypergraph partitioning: Application in VLSI domain," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 7, no. 1, pp. 69–79, Mar. 1999.
- [10] J. N. Kozhaya, S. R. Nassif, and F. N. Najm, "A multigrid-like technique for power grid analysis," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 21, no. 10, pp. 1148–1160, Oct. 2002.
- [11] H. Li, Z. Qi, S. X.-D. Tan, L. Wu, Y. Cai, and X. Hong, "Partitioning-based approach to fast on-chip decoupling capacitor budgeting and minimization," in *Proc. DAC*, San Diego, CA, 2005, pp. 170–175.
- [12] M. Pant, P. Pant, and D. Wills, "On-chip decoupling capacitor optimization using architectural level current signature prediction," in *Proc. IEEE Midwest Symp. Circuits Systems*, Lansing, MI, 2000, pp. 772–775.
- [13] Z. Qi, H. Li, J. Fan, S. X.-D. Tan, Y. Cai, and X. Hong, "On-chip decoupling capacitor budgeting by sequence of linear programming," in *Proc. IEEE ASICON*, Shanghai, China, 2005, pp. 70–73.
- [14] Z. Qi, H. Li, S. X.-D. Tan, L. Wu, Y. Cai, and X. Hong, "Fast decap allocation algorithm for robust on-chip power delivery," in *Proc. ISQED*, San Jose, CA, 2005, pp. 542–547.
- [15] H. F. Qian, S. R. Nassif, and S. S. Sapatnekar, "Random walks in a supply network," in *Proc. DAC*, Anaheim, CA, 2003, pp. 93–98.
- [16] C. K. S. Zhao and K. Roy, "Decoupling capacitance allocation and its application to power-supply noise-aware floorplanning," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 21, no. 1, pp. 81–92, Jan. 2002.
- [17] L. Smith, "Decoupling capacitor calculations for cmos circuits," in *Proc. IEEE Topical Meeting Electrical Performance Electronic Packaging*, Monterey, CA, 1994, pp. 101–105.
- [18] H. Su, S. S. Sapatnekar, and S. R. Nassif, "Optimal decoupling capacitor sizing and placement for standard cell layout designs," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 22, no. 4, pp. 428–436, Apr. 2003.
- [19] J. Vlach and K. Singhal, *Computer Methods for Circuit Analysis and Design*. New York: Van Nostrand Reinhold, 1995.
- [20] J. M. Wang and T. V. Nguyen, "Extended Krylov subspace method for reduced order analysis of linear circuit with multiple sources," in *Proc. DAC*, Los Angeles, CA, 2000, pp. 247–252.
- [21] K. Wang and M. Marek-Sadowska, "On-chip power supply network optimization using multigrid-based technique," in *Proc. DAC*, Anaheim, CA, 2003, pp. 113–118.
- [22] M. Zhao, R. V. Panda, S. S. Sapatnekar, and D. Blaauw, "Hierarchical analysis of power distribution networks," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 21, no. 2, pp. 159–168, Feb. 2002.



Hang Li (S'05) received the B.S. and M.S. degrees in electrical engineering from Beijing Polytechnic University, Beijing, China, and the University of Cincinnati, Cincinnati, OH, in 2000 and 2002, respectively. He is currently working toward the Ph.D. degree at the University of California, Riverside.

He currently co-ops as a Research Staff Member at the R&D Department, Micron Imaging, Micron Technology Inc., San Jose, CA. His research interests include high-performance on-chip power network analysis and optimization and fast thermal circuit simulation.



Jeffrey Fan (S'05) received the B.S. degree in electronics engineering from the National Chiao Tung University, Taiwan, R.O.C., the M.S. degree in electrical engineering from the State University of New York, Buffalo, in 1983 and 1987, respectively, and is currently working toward the Ph.D. degree in electrical engineering at the University of California, Riverside.

From 1988 to 2002, he held various senior technical positions at Western Digital, Emulex Corporation, Adaptec Inc., and Toshiba America. He served as Vice President of Vivavr Technology, Inc., and General Manager/Co-Founder of Musica Technologies, Inc. His research interests include very large scale integration simulation, modeling, and power grid optimization.



Zhenyu Qi (S'04) received the B.S. degree in electrical engineering from Fudan University, Shanghai, China, in 2003 and the M.S. degree in electrical engineering from the University of California, Riverside in 2004, and is currently working toward the Ph.D. degree at the University of California, Riverside.

His research interests include interconnect modeling, high-performance on-chip power/ground network optimization, and fast simulation for RF and thermal circuits.



Sheldon X.-D. Tan (S'96–M'99–SM'06) received the B.S. and M.S. degrees in electrical engineering from Fudan University, Shanghai, China, in 1992 and 1995, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Iowa, Iowa City, in 1999.

From 1995 to 1996, he was a Faculty Member at the Electrical Engineering Department, Fudan University. He is currently an Assistant Professor at the Department of Electrical Engineering, University of California, Riverside. He also coauthored the book

Symbolic Analysis and Reduction of VLSI Circuits (Springer/Kluwer, 2005). His research interests include several aspects of design automation for very large scale integration (VLSI) integrated circuits—modeling, analysis, and optimization of mixed-signal/RF/analog circuits, high-performance and intelligent embedded systems, signal integrity issues in VLSI physical design, high performance power/ground distribution network design and optimization, and VLSI thermal analysis and optimization.

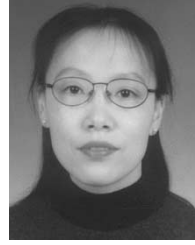
Dr. Tan is a technical program committee member of ISCAS'04, ASP-DAC'05, ISCAS'05, BMAS'05, ASPDAC'06, ISQED'06, and ISCAS'06. He received the Best Paper Award in the 1999 IEEE/ACM Design Automation Conference, the UC Regent's Faculty Fellowship, in 2004, a National Science Foundation CAREER Award, in 2004, and a Best Paper Award Nomination at the 2005 IEEE/ACM Design Automation Conference.



Lifeng Wu (M'95–A'95–M'03) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1984, 1986, and 1990, respectively.

In 1990, he was an Assistant Professor at the Institute of Microelectronics, Tsinghua University. Since 1993, he has been a Research Associate at the University of Washington and the University of California, Berkeley. He joined BTA Technology, in 1995, and Celestry Design Technologies, in 2000. He is currently with Cadence Design Systems, San Jose,

CA. His research interests include analog and mixed-signal circuit simulation, power grid IR drop and electromigration analysis, MOSFET device modeling, hot carrier, and Negative Bias Temperature Instability (NBTI) reliability modeling and simulation.



Yici Cai (M'05) received the B.S. degree in electronic engineering and the M.S. degree in computer science and technology from Tsinghua University, Beijing, China, in 1983 and 1986, respectively.

She is a Full Professor in the Department of Computer Science and Technology, Tsinghua University. Her research interests include signal integrity issues in very large scale integration physical design, power/ground distribution network design and optimization, and high-performance clock network design and optimization.



Xianlong Hong (M'95–SM'95–F'04) received the degree in computational mathematics from Tsinghua University, Beijing, China, in 1964.

In 1964, he joined Tsinghua University, where he is currently a Professor in the Department of Computer Science and Technology. He was a Visiting Scholar at the University of California (UC), Berkeley, from 1991 to 1992, and at UC, Los Angeles, in 1995. As a Technical Leader, he organized and managed to develop three generations of a very large scale integration (VLSI) CAD system, the national

projects in China from 1981 to 1991. He has authored and coauthored over 200 papers and five books around his area. His research interests include physical design for VLSI circuits.

Mr. Hong served as the Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS and as TPC Co-Chair of ASPDAC, in 1999 to 2005. Due to his contribution in developing VLSI CAD systems, he was awarded with more than 15 Science and Technology Achievement Prizes by the Science and Technology Ministry and Education Ministry of China.