Simultaneous Area Minimization and Decaps Insertion for Power Delivery Network Using Adjoint Sensitivity Analysis with IEKS Method

Yu-Min Lee[1], Jeng-Liang Tsai[1] and Charlie Chung-Ping Chen[2]
[1]Department of Electrical and Computer Engineering
University of Wisconsin at Madison
Madison, WI 53706
[2]Institute of Electronics Engineering and Department of Electrical Engineering
National Taiwan University
Taipei 106, Taiwan
yu-min@cae.wisc.edu, jltsai@cae.wisc.edu and cchen@cc.ee.ntu.edu.tw

Abstract— **The soaring clocking frequency and integration density demand robust and stable power delivery to support tens of millions of transistor switching. In this paper, we consider the problem of minimizing the area of wires and decoupling capacitors (decaps) for a power delivery network, subject to the limit on integral of voltage drops. First, we derive the gradients of constraint function without Tellegen's theorem. This greatly simplifies the discuss of adjoint sensitivity analysis. Then, we apply the IEKS method to speed up the sensitivity analysis over 3 times. Finally, this efficient analyzer is incorporated with the state-of-the-art nonlinear programming package, SNOPT, to perform the optimization. Extensive experimental results show that the proposed method can work efficiently for large power delivery networks.**

I. INTRODUCTION

With the ever-increasing clock frequency and the aggressively shrinking feature sizes of high speed electronic circuits, power delivery is becoming a critical design issue. The improper design of power distribution system can degrade the circuit performance, and the reliability. Two basic problems are the narrowing noise margins caused by voltage drops, and the undesirable wear-out of metal wiring caused by electromigration. Given a topology of power delivery network, several techniques may be used for improving the quality of power delivery system: varying widths of wire segments, and adding decoupling capacitors. Wire-sizing has been shown to be an effective way to reduce the power dip/ground bounces as well as improving the electromigration. However, it is too expensive to use wiring sources freely. Consequently, it is necessary to minimize the area of power grid network [1], [2], [3], [4]. Most of the existing methods [1], [2] modeled the network as a resistive mesh with different constant currents consumed by different blocks. The design under constant currents consumption is not reliable with respect to current variations caused by time-variant current waveforms. Those variations can induce higher voltage drops than the expected. This problem can be remedied by over designing the power delivery network. While, the wiring sources will be wasted. Although [3] modeled block currents as random variables to take into account current variations. They still did not consider the dynamic effects, capacitive or inductive, which is significant in high performance circuits. [4] included the dynamic effect and considered the structure of power delivery circuit as a global mesh feeding local trees. They applied PRIMA [5] to calculate the transient adjoint sensitivity over multiple intervals, and a proposed heuristic optimizer to minimize the area.

In this paper, we first model the power delivery network as a lumped RLC equivalent mesh circuit and attach a worst case time-variant current profile at each node. Those current profiles can be estimated by several current extraction methods [6], [7]. Then, we use the PWL (piece wise linear) functions to approximate those profiles. After that, we use the integral of voltage drop below a specific noise margin [8] as the noise metric function for each node. Later on, we use the sum of metric function at each node as the measure function, and develop an efficient adjoint sensitivity analyzer with suitable model order reduction techniques to calculate the gradients of measure function with respect to each wire width, and each decoupling capacitor. Finally, we incorporate the above sensitivity calculation method with the state-of-the-art nonlinear programming algorithm, SQP (sequential quadratic programming) with SNOPT [9], to minimize the occupied area of power delivery network.

The rest of the paper is organized as follows. First, the equivalent circuit model we use for the power delivery network will be introduced, and the adjoint sensitivity of voltage drop integral with model order reduction techniques will be derived in Section II. Then, the formulation of the tuning problem and optimization method will be presented in Section III. Finally, the numerical experiments and conclusion will be given in Section IV, and V.

II. POWER DELIVERY CIRCUIT AND ITS SENSITIVITY COMPUTATION

The power delivery structure is represented by an equivalent circuit shown in Figure 1. An independent time-variant PWL waveform is attached at each node to represent the drawn current of cell, and each wire segment is modeled by a resistor and inductor connected in series and a ground capacitor. The RLC parameters of each wire are given by

$$R_s = \rho l_s / w_s \quad (1)$$

$$C_s = (\beta w_s + \alpha) l_s \quad (2)$$

$$L_s = \gamma l_s / w_s \quad (3)$$

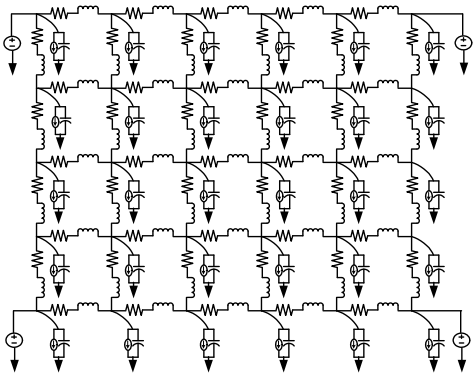

Fig. 1. Equivalent Circuit Model of Power Delivery Network

where l_s and w_s are the length and width of each wire segment, and ρ, β, α, γ are the sheet resistance per square, sheet capacitance per square, fringe capacitance per unit length and inductance per square of the metal layer.

The behavior of such a system can be expressed by the MNA (modified nodal analysis) [10] formulation as a first ordinary differential equation,

$$\mathbf{G}(p)v(t,p) + \mathbf{C}(p)\dot{v}(t,p) \quad = \quad \mathbf{B}u(t), \tag{4}$$

where $v(t,p)$ is the vector of variables, $u(t)$ denotes a vector of port voltage sources and internal current sources which are represented by PWL functions, $\mathbf{G}(p)$ is the conductance matrix, $\mathbf{C}(p)$ is the susceptance matrix, $\mathbf{B}$ is the input selector matrix mapping the sources to the internal states, and p is the vector of tunable parameters which are the widths of wire segments or decoupling capacitors in our case. Circuit equations as shown in Equation (4) can be transformed to the $s-$ domain by Laplace transformation as

$$\mathbf{G}(p)\mathbf{v}(s,p) + s\mathbf{C}(p)\mathbf{v}(s,p) \quad = \quad \mathbf{Bu}(s) + \mathbf{C}v(0), \tag{5}$$

where **v(s,p)** and $\mathbf{u}(s)$ are the Laplace transform of $v(t,p)$ and $u(t)$, and $v(0)$ is the initial condition of $v(t,p)$.

The integral of voltage drop below a specified noise margin was first introduced in [8] and was proved to be an efficient noise metric for the performance of each node in the power distribution network [11]. This integral according to Figure 2 can be described as

$$\begin{aligned} c_i(p) &= \int_0^T \max\{NM_H - v_i(t,p), 0\}dt \\ &= \int_0^T g_i(t)\left(NM_H - v_i(t,p)\right)dt, \end{aligned} \tag{6}$$

where $v_i(t,p)$ is voltage drop at node i, and $g_i(t)$ is an unit pulse within time interval $[t_{s_i}, t_{e_i}]$.

The most critical node in the power delivery networks is defined as the largest voltage drop integral over desired time period T. Instead of using the voltage drop integral of most critical node as a metric, we choose the measure function to be the sum of voltage drop integral at each node.

$$c(p) \quad = \quad \sum_i c_i(p) \tag{7}$$

Therefore, we emphasize more on the average global effect rather than only the effect of most critical node. The goal of our optimization problem is to use the minimum amount of area for wiring power distribution network while the sum of voltage drop integral of whole circuit is less than or equal to zero.

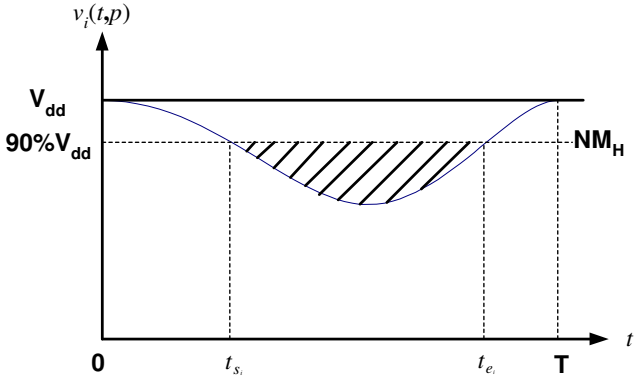

Fig. 2. Voltage Drop at Node i

A. Integral Sensitivity Derivation without Tellegen's Theorem

During the procedure of optimization, one is sometimes interested in the sensitivity of the performance function with respect to many parameter values. Adjoint sensitivity analysis [10], [12] is a well known efficient technique to calculate the sensitivities of one performance function with respect to many parameter values is required. The basis for adjoint analysis comes from Tellegen's theorem [13]. We are going to introduce a more easy way to analyze the sensitivity without using Tellegen's theorem.

To illustrate the procedure of sensitivity calculation for function $c(p)$, we first define a new function $c(t,p)$ as

$$\begin{aligned} c(t,p) &= \sum_i \int_0^t g_i(\tau)\left(NM_H - v_i(\tau,p)\right)d\tau, \\ &= \sum_i \int_0^t h_i(t-\tau)\left(NM_H - v_i(\tau,p)\right)d\tau, \end{aligned} \tag{8}$$

where $h_i(t) = g_i(\tau - t)$ for each i (for example, $h_i(t)$ is a unit plus within time interval $[T - t_{s_i}, T - t_{e_i}]$ with respect to Figure 2). Equation (8) can be expressed as the following vector form

$$c(t,p) \quad = \quad \int_0^t h^T(t-\tau)\left(\mathbf{NM_H} - v(\tau,p)\right)d\tau. \tag{9}$$

where $h(t)$ is a functional vector with each entry i being $h_i(t)$, and $\mathbf{NM_H}$ is a vector with each entry equal to NM_H. Equation (9) is a convolution-representation of $c(t,p)$. After taking Laplace transformation on its both sides, and utilizing Equation (5), we have

$$\mathbf{c}(s,p) \quad = \quad \mathbf{h}^T(s)\left(\frac{1}{s}\mathbf{NM_H} - \mathbf{A}^{-1}(s,p)\mathbf{b}(s)\right), \tag{10}$$

where $\mathbf{c}(s,p)$ and $\mathbf{h}(s)$ are the Laplace transform of $c(t,p)$ and $h(t)$, $\mathbf{A}(s,p) = \mathbf{G}(p) + s\mathbf{C}(p)$, and $\mathbf{b}(s) = \mathbf{Bu}(s) + \mathbf{C}v(0)$. Therefore,

$$\begin{aligned} \frac{\partial \mathbf{c}}{\partial p_k} &= \mathbf{h}^T\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial p_k}\mathbf{v} \\ &= {\mathbf{v}_a}^T\frac{\partial \mathbf{A}}{\partial p_k}\mathbf{v}, \end{aligned} \tag{11}$$

where $\mathbf{v}$ is the solution of original MNA equations and $\mathbf{v}_a$ is the solution of adjoint MNA equations in $s-$ domain,

$$\mathbf{A}\mathbf{v} = \mathbf{b} \quad (12)$$

$$\mathbf{A}^T\mathbf{v}_a = \mathbf{h} \quad (13)$$

Finally, the sensitivity of $c(p)$ with respect to an arbitrary parameter p_k is

$$\frac{\partial c(p)}{\partial p_k} = \left.\frac{\partial c(t,p)}{\partial p_k}\right|_{t=T} = \int_0^T {v_a}^T(T-\tau)\left[\frac{\partial \mathbf{G}}{\partial p_k}v(\tau) + \frac{\partial \mathbf{C}}{\partial p_k}\dot{v}(\tau)\right]d\tau \quad (14)$$

In order to calculate $\partial c(p)/\partial p_k$, we need to know the waveforms of $v_a(t)$ and $v(t)$ which can be done by applying trapezoidal integration approximation of Equation (12)–(13) in the time domain and solving them easily only two forward/backward substitutions at each time step. [14] proposed a Preconditioned Conjugate Gradient iterative method to efficiently solve MNA equations. However, due to the large size of power delivery network and tremendous number of MNA solving needed during the procedure of optimization, it still consumes a lot of computational times.

B. Integral Sensitivity Computation with Model Order Reduction Techniques

Model order reduction techniques have shown to be a very efficient way to speed up the circuit analysis [10], and have been widely studied and improved over the last decade [15], [16], [5], [17]. Starting from AWE (Asymptotic Waveform Evaluation) [15] to PRIMA [5] (Passive Reduction Interconnect Macromodeling Algorithm), model order reduction techniques have been successfully extended to consider inductance effects in a reasonable accuracy. Later, [17] developed the EKS (Extended Krylov Subspace) method to simulate large scale power delivery circuits with many PWL current sources. To resolve the source waveform modeling issues, EKS has to perform the moment shifting procedure to recover the proper moments. Recently, [18] proposed the IEKS (improved-EKS) method such that it no longer needs to perform moment shifting for source waveform modeling. The major advantage of EKS/IEKS method is their runtime not proportional to the number of independent sources. Since the power distribution network contains lots of current sources, our sensitivity computation is coupled with the IEKS-based order reduction approach so that it can handle the large-scale power delivery circuit.

Given a power delivery circuit with the system Equation (5), we apply IEKS method to compute the orthonormal basis $\mathbf{X}$ of its extended Krylov subspace. Then, we construct its order-reduced model by projecting the original system $(\mathbf{G}, \mathbf{C}, \mathbf{B}, \mathbf{v})$ onto this subspace via congruent transformation, $\widehat{\mathbf{G}} = \mathbf{X}^T\mathbf{G}\mathbf{X}$, $\widehat{\mathbf{C}} = \mathbf{X}^T\mathbf{C}\mathbf{X}$, $\widehat{\mathbf{B}} = \mathbf{X}^T\mathbf{B}$, and $\widehat{\mathbf{v}} = \mathbf{X}^T\mathbf{v}$. The dimension of this new system $(\widehat{\mathbf{G}}, \widehat{\mathbf{C}}, \widehat{\mathbf{B}}, \widehat{\mathbf{v}})$ is reduced because the rank of $\mathbf{X}$ is much smaller than the original matrix $\mathbf{A}$. Therefore, the runtime much less than the original circuit. After that, we set up the system equations of reduced circuit and utilize the fast simulation method in [14] to get the waveform of $\widehat{v}(t)$. The $\widehat{v}(t)$ is projected back to the original space to provide the approximate solution, $v(t) \approx \mathbf{X}\widehat{v}(t)$. Details of the IEKS reduction procedure could be found in [18]. Finally, we use $v(t)$ to construct the excitation, $g_i(t)$, at each node i of the original system.

We, then repeat the above procedure to compute the solution, $v_a(t)$, of adjoint circuit, and plug $v(t)$ and $v_a(t)$ back to Equation (14) to get the sensitivity of $c(p)$ with respect to each tunable parameter p_k.

III. OPTIMIZATION

The problem of minimizing the area of power delivery network by varying the widths of wire segments can be formulated as

$$\begin{aligned} \text{minimize} \quad & \sum_i l_i w_i \\ \text{subject to} \quad & c(w) \le 0, \end{aligned} \quad (15)$$

where l_i and w_i are the length and width of wire segment i, w is the vector of wire widths, and $c(w)$ represents the sum of voltage drop integral.

The optimization engine is based on the state-of-the-art nonlinear programming technique SQP (sequential quadratic programming) method with SNOPT [9], [19], and our adjoint sensitivity analyzer. During the procedure of optimization, the analyzer continuously simulates the network, computes its sensitivities by using the method presented in Section II-B, and provides those sensitivities to the SNOPT-based optimizer.

SNOPT employs a limited memory quasi-Newton approximation [20] to the Hessian of the Lagrangian and augmented Lagrangian merit function. It uses an active set approach with only first order information. A modified Lagrangian function is employed where the algorithm finds the stationary point to the Lagrangian by solving a sequence of quadratic approximations. Please refer to [9], [19] for the detail description of SNOPT.

The above optimization engine can be modified to include the decoupling capacitors as

$$\begin{aligned} \text{minimize} \quad & \mu\sum_i l_i w_i + \nu\sum_j \varpi_{C_j} \\ \text{subject to} \quad & c(w, \varpi) \le 0, \end{aligned} \quad (16)$$

where ϖ_{C_j} is the area of decoupling capacitor C_j, ϖ is a vector of area of decoupling capacitors, and μ, ν are weighting factors. A tunable decoupling capacitor is attached at each mesh node, and is initially set to zero. We can apply the same technique in Section II-B to calculate the gradients of $c(w, \varpi)$ with respect to these decoupling capacitors and wire segments. Then, the SNOPT-based optimizer is utilized to find the minimum weighted sum of the area of wires and the area of all decoupling capacitors.

IV. EXPERIMENTAL RESULTS

We implement the proposed sensitivity analysis method in C++ language, and apply the SNOPT as the kernel of our optimization engine. All results are performed on a PC with a 1.4GHz Pentium IV processor. The typical parameters for each wire segment are $\rho = 0.022\Omega$, $\beta = 0.018fF/\mu m^2$,

$\alpha = 0.040fF/\mu m$, and $\gamma = 1.26pH$. The supply voltage is 1 volt, and NM_H is equal to 0.9 volt.

Table I lists the runtime of one full integral sensitivity analysis implemented by the IEKS method or an efficient MNA solver [14]. The topology of each circuit is mesh, and reduction order of IEKS method is 14 for all of them. It shows that integral sensitivity computation based on model order reduction technique is about 3 times faster than based on the MNA solver [14]. The tendency that the speed up increases with larger circuit size is shown.

The error distribution of integral sensitivity analysis based on the IEKS method is illustrated in Figure 3. The mesh size is 81×81 (12960 wire segments), and the order of reduction is 14. It demonstrates that our method is quite accurate. The amplitude of maximum error is less than 5%, and the errors for 96% wires are within $\pm 1\%$.

Table II represents the results of our optimization engine for four different mesh circuits without decoupling capacitors. The size of these circuits are from 26×41 nodes to 61×66 nodes. The number of nodes, number of wires, and the minimum wire area of each circuit are listed. The CPU times are listed in the last column.

Circuit	# of Wires	IEKS (s)	MNA Solver (s)	Speedup (X)
m 10 × 10	180	0.201	0.180	0.90
m 30 ×30	1740	1.563	2.834	1.81
m 70 × 70	9660	9.183	23.464	2.56
m 90 × 90	16020	15.813	42.531	2.69
m 100× 100	19800	19.228	55.740	2.90
m 200× 200	79600	89.930	277.959	3.09

TABLE I

RUNTIME COMPARISON OF A FULL INTEGRAL ADJOINT SENSITIVITY ANALYSIS BETWEEN IEKS METHOD BASED AND AN EFFICIENT MNA SOLVER BASED

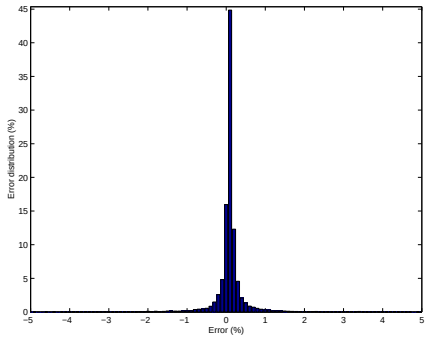

Fig. 3. Error Distribution of Integral Adjoint Sensitivity Analysis Based on IEKS Method

V. CONCLUSION

An easy way to understand and derive the formula of adjoint integral sensitivity for the power delivery network design without Tellegen's Theorem is introduced. An efficient and fast method of analyzing the adjoint integral sensitivity with model order reduction techniques is developed. Numerical results show that the proposed method can ease the computational load with very small error.

This fast gradient calculating method has been combined with a nonlinear optimizer, SNOPT, to efficiently optimize the power delivery network. Although the experimental results do not include the decoupling capacitors, they can be easily added into our optimization engine.

Circuit	# of Nodes	# of Wires	Wire Area (cm^2)	CPU Time (hrs)
m 26 × 41	1066	2065	0.00545	0.13
m 41 × 51	2091	4090	0.01040	1.26
m 51 × 51	2601	5100	0.01272	2.13
m 61 × 66	4026	7925	0.02036	5.81

TABLE II

RESULTS OF OPTIMIZATION

References

[1] S. Chowdhury and M. A. Breuer. Optimum design of ic power/ground nets subject to reliability constraints. In *IEEE Trans. on Computer-Aided-Design*, volume 7(7), pages 787–796, 1988.

[2] X. tan, C. J. Richard Shi, D. Lungeanu, L. Yuan, and J. Lee. Reliability-constrainted area optimization of vlsi power/ground networks via sequence of linear programming. In *DAC*, pages 78–83, 1999.

[3] S. Boyd, L. Vandenberghe, and A. El Gamal. Design of robust global power and ground networks. In *ISPD*, 2001.

[4] Haihua Su, Kaushik Gala, and Sachin S. Sapatnekar. Fast analysis and optimization of power/ground networks. In *ICCAD*, pages 477–480, 2000.

[5] A. Odabasioglu, Mustafa Celik, and L.T. Pillage. Prima: Passive reduced-order interconnect macromodeling algorithm. In *ICCAD*, pages 58–65, 1997.

[6] A. Krstic and K. Cheng. Vector generation for maximum instantaneous current through supply lines for cmos circuits. In *DAC*, pages 383–388, 1997.

[7] S. Bobba and I. N. Hajj. Estimation of maximum current envelope for power bus analysis and design. In *ISPD*, pages 141–146, 1998.

[8] A. R. Conn, R. A. Haring, and C. Visweswariah. Noise considerations in circuit optimization. In *ICCAD*, pages 220–227, San Jose, CA, 1998.

[9] P. E. Gill, W. Murray, and M. A. Saunders. User's guide for snopt 5.3: A fortran package for large-scale nonlinear programming. Technical report, System Optimization Laboratory, Stanford University, Stanford, CA, 1997.

[10] L. T. Pillage, R. A. Rohrer, and C. Visweswariah. *Electronic Circuit and System Simulation*. McGraw-Hill Book Co., 1995.

[11] H. Su, S. S. Sapatnekar, and S. R. Nassif. Optimal decoupling capacitor sizing and placement for standard-cell layout designs. In *IEEE Trans. on Computer-Aided-Design*, volume 22(4), pages 428–436, 2003.

[12] S. W. Director and R. A. Rohrer. The generalized adjoint network and network sensitivities. In *IEEE Trans. on Circuit Theory*, volume 16(3), pages 318–323, August 1969.

[13] C. A. Desoer and E. S. Kuh. *Basic Circuit Theorey*, volume 2, pages 292–300. McGraw-Hill Book Co., 1967.

[14] Tsung hao Chen and Charlie Chung-Ping Chen. Efficient large-scale power grid analysis based on preconditioned krylov-subspace iterative methods. In *DAC*, pages 559–562, 2001.

[15] L. Pillage and R. A. Rohrer. Asymptotic waveorm evaluation for timing analysis. In *IEEE Trans. on Computer-Aided Design*, volume 9, pages 352–366, April 1990.

[16] R. W. Freund and P. Feldman. Reduced-order modeling of large passive linear circuits by menas of the sypvl algorithm. In *ICCAD*, pages 280–287, 1996.

[17] Janet M. Wang and Tuyen V. Nguyen. Extended krylov method for reduced order analysis of linear circuits with multiple sources. In *DAC*, pages 247–252, 2000.

[18] Yahong Cao, Yu-Min Lee, Tsung-Hao Chen, and Charlie Chung-Ping Chen. Hiprime: Hierarchical and passivity reserved interconnect macromodeling engine for rlkc power delivery. In *DAC*, pages 379–384, 2002.

[19] P. E. Gill, W. Murray, and M. A. Saunders. Snopt: A sqp algorithm for large-scale constrained optimization. Technical report, Department of Mathematics, University of California, San Diego, La Jolla, CA, 1997.

[20] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Verlag, Heidelberg, Berlin, New York, 1999.