Voltage Drop Reduction For On-chip Power Delivery Considering Leakage Current Variations

Jeffrey Fan[†], Ning Mi[‡], Sheldon X.-D. Tan[‡],
[†]Department of ECE, Florida International University, Miami, FL 33174, USA
[‡]Department of EE, University of California, Riverside, CA 92521, USA *

Abstract

In this paper, we propose a novel on-chip voltage drop reduction technique for on-chip power delivery networks of VLSI systems in the presence of variational leakage current sources. The new method inserts decoupling capacitors (decaps) into the power grid networks to reduce the voltage fluctuation. The optimization is based on sensitivity-based conjugate gradient method and sequence of linear programming approach. Different from existing power grid noise reduction methods, the new approach considers the impacts of inter-die and intra-die variational leakage current sources due to unavoidable process variability during the decap optimization process for the first time. Leakage currents, which although are static in nature typically, can still add to the total voltage drops and dynamic voltage reduction thus must consider the leakage-induced voltage variations. The proposed algorithm exploits the relative constant variations for different decap configurations of power grid circuits to speed up the statistical optimization process. Decaps can be inserted in such a way that the resulting circuits have much higher probability to meet the voltage drop constraints in the presence of leakage current variations. Experimental results demonstrate the effectiveness of the proposed approach and show that the new method has 100X to 1,000X of speedup over the Monte Carlo based statistical decap optimization method.

1 Introduction

Due to increasing integration density and soaring clock frequency, reliable on-chip power supply becomes a critical concern for VLSI chip design, especially in the deep submicron arena. To retain signal integrity in power grid (P/G) network, which supplies power from the power/ground pads to all the modules on a chip, extra careful design effort is required to reduce the voltage noise in the on-chip power supply networks. However, excessive voltage variations, including voltage drops and ground bounces, would have an adverse impact on chip performance and reliability, since they not only degrade the already tight noise margin in today's VLSI circuits, but also increase gate delay, which may cause false logic switching, and sometimes even lead to logic failure. In practice, most designs now require voltage drop to be confined to certain percentage of the nominal supply voltage [2].

On top of the reduced noise margin, process-induced variations in nanometer VLSI systems make voltage *IR* drop more susceptible in system performance. One of the major variations comes from the leakage current variations as leakage currents are exponential functions of the threshold voltages, which typically are random variables due to the variations of channel length and gate oxide thickness. As a result, the leakage currents are very sensitive to those parameter variations. It was shown in [11] that leakage variations for 90*nm* can be 20 *times*. Based on the ITRS 2005 [1], the leakage power accounts for more than 60% at 45*nm* technology. This has many consequences in chip design, especially in the design of power grid network. In this paper, we model these leakage current sources as log-normal distribution due to threshold voltage variation, which is model as Gaussian (normal) distribution. [9] [13] Typically, the

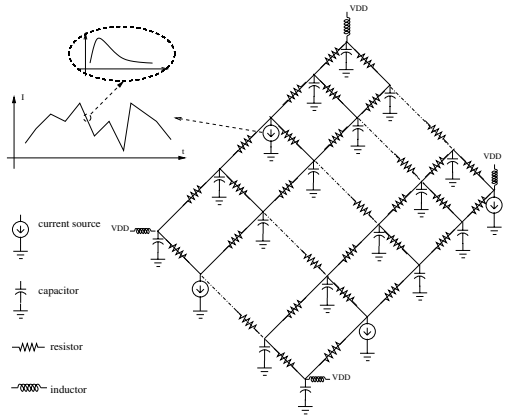

Figure 1. Illustration of decaps in a power grid network.

process-induced variations can be classified into inter-die (die-to-die) variations and intra-die (within die) variations. In inter-die variations, all the parameters of variations are correlated. The worst case may be easily calculated by setting the parameters to their top range, i.e. mean plus three

*This work is supported in part by NSF CAREER Award under grant CCF-0448534, and NSF grant OISE-0623038.

standard deviations. The intra-die variations consist of two components, i.e. layout-dependent deterministic components and stochastic (random) components. To consider the correlations among random variables, a hierarchical structure of variables is used in our work, which models certain variables as global variables (inter-die), and divides variational sources into several sections (intra-die) based on the geographical location of the sources. Thus, the total variations are subject to both inter-die and intra-die variations.

A commonly-used, yet effective way to reduce noise in the package is to add decoupling capacitors (decaps) into the power grid networks [5, 14]. For VLSI design, the inductive wire impedance can no longer be ignored at high frequencies, and the on-chip noise becomes more profound as inductance scales poorly with wire sizing. As a result, on-chip decaps are indispensable for robust on-chip power supply [15, 4, 17]. On the other hand, on-chip decaps are usually manufactured as gate capacitance of transistors. As the supply voltage continues scaling, leakage currents due to reduced threshold voltage and dielectric leakage prevent excessive use of decaps [3]. Thus, the economic use of on-chip decaps is relatively important, especially in the presence of leakage current variations.

Note that leakage current typically is static current. Nevertheless, this leakage current becomes more dynamic as several leakage reduction techniques, such as sleep transistor, adaptive body bias, and input vector method, have been proposed recently. Adding decap only reduces the dynamic voltage fluctuation. However, the voltage drop induced by leakage current will manifest itself as background noise. This type of noise increases the dynamic range of voltage fluctuation in reference to the nominal absolute DC values. This effect has huge impact on the circuit delay and operation. As a result, in order to limit the total voltage drop under the same given threshold voltage, more decaps will need to be allocated. Hence, decap allocation can still be used to mitigate the leakage induced voltage drops as well as the voltage drops caused by dynamic powers of the circuits.

The contributions of the paper are summarized as follows: (1) We first show that in the presence of leakage current variations, traditional deterministic decap optimization is unable to sufficiently remove all the *IR* drop violations without statistics-based approach. (2) We adopt two efficient statistical decap allocation algorithms, based on improved conjugate gradient optimization and sequence of linear programming, to show that our approach may apply to any known decap optimization algorithms. (3) We will show in the experiment that the calculated standard deviation based on Hermite polynomials is insensitive to the added decaps in each iteration of the decap optimization, thus dramatically improve the speed of solving decaps. (4) In consideration of correlations in inter-die and intra-die variations, the hierarchical structure [10] of random variables are used in the decap optimization process.

2 Statistical Decap Allocation Problem

Fig. 1 illustrates the model of power grid network considering process variations. The deterministic component is modeled as piecewise linear (PWL) current source including certain nominal value of DC leakage current. The total leakage currents are the effects contributed from the deterministic leakage current and the stochastic component, which is modeled as normalized lognormal distribution in consideration of process variations.

Before finding the solution for decap optimization, we need to formulate power grid networks considering process variations mathematically. Let Θ denote the process sampling space. Let $\theta \in \Theta$, $\xi_i : \theta \rightarrow R$ denote a normalized Gaussian variable, and $\xi(\theta) = [\xi_1(\theta), ..., \xi_n(\theta)]$ is a vector of n Gaussian variables. Therefore, given the process variations, the Modified Nodal Analysis (MNA) of the power grid network becomes

$$Gv(t,\xi(\theta)) + C\frac{dv(t,\xi(\theta))}{dt} = I(t,\xi(\theta)) \qquad (1)$$

Note that the input current vector, $I(t,\xi(\theta))$, consists of both deterministic and stochastic components.

The problem we need to solve is to efficiently calculate the mean and variances of decap values (budgets) due to variational current sources. For given current inputs, decap values are computed to prevent the *IR* drops from exceeding the user-defined thresholds. With those statistical decap information, we may add the decaps in a consecutive manner, for instance means plus three standard deviations, to ensure that the *IR* drop constraints are satisfied with higher possibility. A straightforward method is Monte Carlo (MC) based sampling approach. It means that we randomly generate $I(t,\xi(\theta))$, which is based on the lognormal distribution, and solve for the minimum decaps by using existing optimization approach for each sampled circuit. Obviously, it is not scalable because MC is computationally expensive. In consideration of stochastic techniques, the error associated to MC simulation roughly scales with $1/\sqrt{M}$, where M is the number of samplings or trials. Thus, $10,000$ trials are necessary to guarantee 1% of accuracy [16].

3 New Statistical Decap Allocation Algorithm

In this section, we present the new statistical decap allocation algorithms. The main idea of the new algorithm is to decompose the optimization problem into sequential two parts: (1) computing the statistical node voltage variations using fast statistical spectrum method. (2) determining the variational decap values based on the variational node voltages during decap allocation process. The novelty of the new method is based on the observation that node voltage variations due to leakage current variations do not change significantly with different decap values. As a result, we can separate the variational node voltage computations and decap optimization process to save significant computing costs.

3.1 Variations in voltages and violation areas

In this subsection, we first discuss the variations on the *IR* drop or equivalently *ground bounces* at the nodes of the

power grid network because of the variations of the leakage current sources. Then we show how it can be efficiently computed.

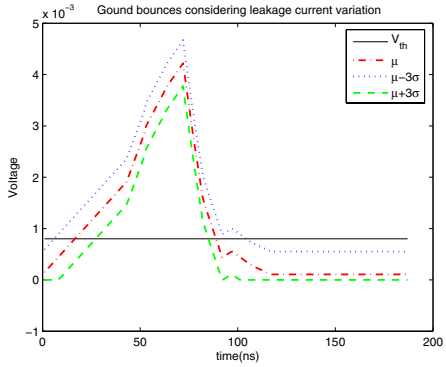

Figure 2. Illustration of ground bounce at given node considering variations.

Fig. 2 illustrates the ground bounces at a given node considering the variations of leakage current sources. The horizontal line V_{th} denotes the *threshold voltage*. The middle dash line, μ, is the deterministic voltage at the arbitrary node. The upper and lower dash lines (called *upper bound* and *lower bound* respectively in this paper) indicate the voltages when three *standard deviations* (σ) for certain distribution are added or deducted from the *mean* (μ) or deterministic voltage.

One may easily observe that the violation areas also follow certain distribution because the voltages are moved up and down around the deterministic node voltages. However, the node voltages should never go negative in calculating violation areas in consideration of lower bound voltages. As a result, the traditional deterministic decap budgets calculated by Improved Conjugate Gradient (CG) [12] or Sequence of Linear Programming (SLP) [6] are unable to sufficiently eliminate all the violation areas at the nodes in the presence of variations of leakage current sources.

Assume that we are able to estimate the mean and standard deviation or variance (to be discussed further later) of the voltage at any given node, then the violation areas may be obtained for the given threshold voltage specified by designers. Once we know the "statistics" in terms of mean and variance of the violation areas, the "statistics" of the decap budgets can approximately obtained by using existing decap optimization approaches, such as CG or SLP in a deterministic way.

3.2 Relative "constant" node voltage deviations

One important observation in this work is that the standard deviations of node voltages are almost unchanged with added decaps in each iteration of CG or SLP optimization. This means that we do not need to solve for all the node voltage distributions in our proposed approach for each optimization iteration. We only need to solve the distribution in terms of standard deviations of the node voltages at the beginning, and use the same results for all the subsequent iterations during decap optimization. This feature dramatically reduces the time in solving the decap budgets with our proposed approach. Fig. 3 shows the relationship between the node voltages in one standard deviation calculated by Hermite polynomial approach (to be discussed) verse added decap budgets in each iteration of CG or SLP optimization for one particular testing case. One may notice that the node voltage remains almost unchanged for various values of standard deviation. We find this is true for all the testing cases we have.

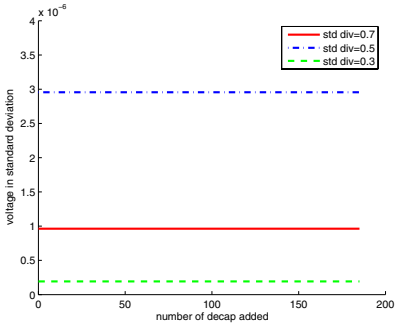

Figure 3. Relationship between voltage in standard deviation and added number of decaps.

The intuitive reason for this phenomenon is that leakage currents are DC-like currents and their induced voltage drops do not change significantly in the transient waveforms. As a result, decap values, which will change the dynamic voltage, do not change the leakage-currents which induce node voltage variations.

3.3 Variations in leakage current sources

In this subsection, we present the approach for modeling the leakage current as lognormal distribution by using Hermite PCs [7].

Let $g(\xi)$ be the Gaussian random variable, denoting threshold voltage or device channel length. Let $l(\xi)$ be the random variable obtained by taking the exponential of $g(\xi)$

$$l(\xi) = e^{g(\xi)},\ g(\xi) = ln(l(\xi)) \tag{2}$$

Also, for the MOSFET device leakage current,

$$I_{off} = cI_l(V_{th}) = ce^{-V_{th}} \tag{3}$$

where the leakage component $I_l(V_{th})$ is a lognormal random variable and c is a constant. Let the *mean* and *variance* of $g(\xi)$ be μ_g and σ_g^2 respectively, then the *mean* and *variance* of $l(\xi)$ are derived as

$$\mu_l = e^{(\mu_g + \frac{\sigma_g^2}{2})} \tag{4}$$

$$\sigma_l^2 = e^{(2\mu_g + \sigma_g^2)}[e^{\sigma_g^2} - 1] \tag{5}$$

respectively. For a general Gaussian variable $g(\xi)$, it can be represented as

$$g(\xi) = \sum_{i=0}^{n} \xi_i g_i \tag{6}$$

where ξ_i are orthonormal Gaussian variables, which means that $< \xi_i, \xi_j >= \delta_{ij}$, $< \xi_i >= 0$ and $\xi_0 = 1$. Note that such form can always be obtained by using Karhunen-Loeve orthogonal expansion method [8]. In our problem, we need to represent the lognormal random variable, $l(\xi)$, by using the Hermite PC expansion form:

$$l(\xi) = \sum_{k=0}^{P} l_k H_k^n(\xi) \tag{7}$$

where $l_0 = \exp[\mu_g + \frac{\sigma_g^2}{2}]$. To find the coefficients, we can apply Galerkin method [7] on $l(\xi)$. Therefore, we have

$$l_k(t) = \frac{< l(t,\xi), H_k(\xi) >}{< H_k^2(\xi) >}, \ \forall k \in \{0,...,P\}. \tag{8}$$

Thus, the lognormal process can then be written as

$$l(\xi) = l_0(1 + \sum_{i=1}^{n} \xi_i g_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{(\xi_i\xi_j - \delta_{ij})}{< (\xi_i\xi_j - \delta_{ij})^2 >} g_i g_j + ...) \tag{9}$$

where g_i is defined in eq. (6). Combining eq. (1) and (9), the mean and variance of node voltages, $v(t)$, can be obtained after some math calculations [13].

Finally, let's consider the correlated multiple random variables. For illustration purpose, as shown in Fig 4, the variations are divided into three layers due to the sources of variations [10]. The global variable (top layer, denoted as level A) is modeled as inter-die variation, such as lot-to-lot, wafer-to-wafer, and die-to-die variations. The intra-die variations, middle layer, denoted as level B and bottom layer as level C, are divided into several sections in the power grid networks. Each section may have different variation in terms of variance or standard deviation. Thus, the correlations among random variables are modeled as hierarchical structure, in such a way that the total variations of any given node are subject to the variations in Level A, B, and C respectively. Similarly, by following the proposed hierarchical structure, one may easily consider the correlations by constructing any given number of variational sources into several layers, and obtain total variation subject to the variations in each layer.

4 Experimental Results

This section describes the simulation results of circuits with lognormal leakage current distributions for a number of power grid networks. All the proposed methods have been implemented in Matlab, Perl and C/C++. The experimental results are carried out in a Linux system with dual Intel Xeon CPUs with 3.06Ghz and 1GB memory.

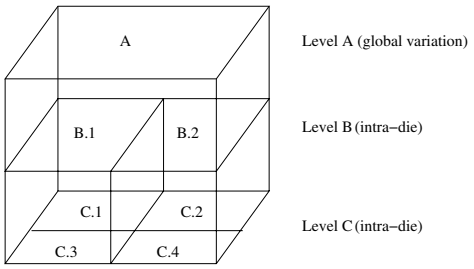

Figure 4. Hierarchical structure of random variables.

4.1 Decap optimization considering one global random variable

Firstly, the proposed method is applied to power grid optimization by adding decoupling capacitors (decaps) considering one global variable variation (*inter-die* variation). The assumption is that the nominal mean value (DC) of leakage current is set to be 300 μA in this experiment. Both the *Conjugate Gradient* (CG) and *Sequence of Linear Programming* (SLP) are used as the optimization engines in our test. Additionally, the leakage currents caused by added decaps are neglected in our experiments because they are relative small after optimization in comparison to the ones caused in the power grid networks.

In Monte Carlo sampling, we first multiply sampled value from normalized lognormal distribution with nominal DC leakage currents, then adding this value into deterministic PWL current sources as one trial of the stochastic leakage current values. Given the threshold voltage specified by designers, we may calculate the *violation areas* of all the nodes in each sampled circuit. Based on the violation areas, both CG approach and SLP method are applied to solve for decap budgets. The Monte Carlo results of the distribution of decap budgets by using CG and SLP for one random variable, as global inter-die variation, are shown in Fig. 5 and Fig. 6. The CG and SLP are both performed for 4000 trials for lognormal leakage current sources with the same standard deviation ($\sigma_{global} = 0.1$). The *x-axis* indicates the decap budget in terms of $farads$ in each trial (occurrence) and *y-axis* denotes the number of occurrences with total trial size of 4000. Next, the Hermite PC approach is performed to calculate the statistics of node voltages before solving for decap optimization by using CG and SLP. The second order of Hermite PC is adopted in calculating the standard deviations of the node voltages in both CG and SLP. The results are marked in the figures. In the figure, the meanings of three vertical lines are as follows: the middle vertical line indicates the deterministic decap value without variations (μ in the figure); the right line denotes the decap budget for the circuit when the node voltages are raised by three standard deviations calculated by Hermite polynomial expansion (the upper bound, $\mu + 3\sigma$); and the left line indicates the decap budget for the circuit when the node voltages are lower by three standard deviations (the lower bound, $\mu - 3\sigma$).

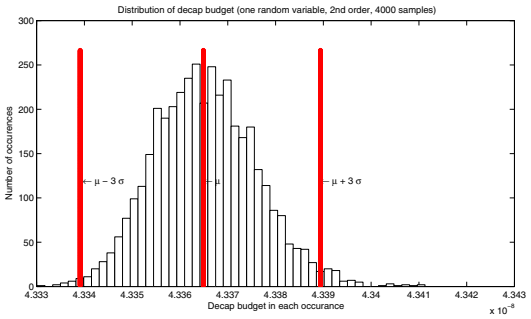

Figure 5. Distribution of decap budget with one variable using CG algorithm.

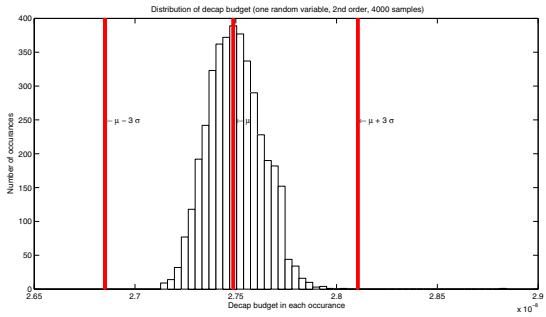

Figure 6. Distribution of decap budget with one variable using SLP algorithm.

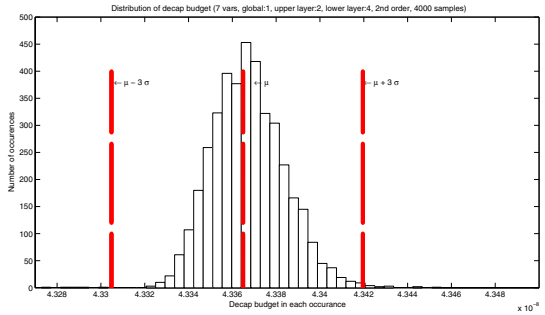

Figure 7. Distribution of decap budget with 7 variables using CG algorithm.

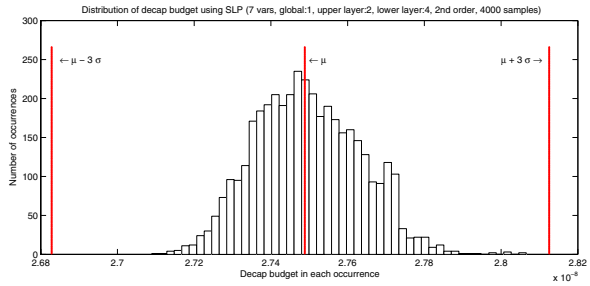

Figure 8. Distribution of decap budget with 7 variables using SLP algorithm.

4.2 Decap optimization considering multiple variables with correlations

The consideration of correlations among random variables are modeled as three layers of hierarchical structure in the experiment as shown in Fig. 4. In CG method, the standard deviation of the global variable (the highest layer, Level A) is assumed to be $\sigma_A = 0.1$ in the experiment. The standard deviations for Level B and C are given as $\sigma_B = 0.1, 0.08$ and $\sigma_C = 0.1, 0.2, 0.16, 0.18$ respectively. When using SLP method, $\sigma_A = 0.1$, $\sigma_B = 0.05, 0.01$ and $\sigma_C = 0.01, 0.01, 0.01, 0.01$ respectively. Fig. 7 and Fig. 8 show the results based on hierarchical structure of seven variables by solving decap budgets in terms of $farads$ with CG and SLP algorithms. The meanings of three vertical lines are the same as in the figure for one global variable, i.e. lines for deterministic, upper bound, and lower bound.

Furthermore, Table 1 shows the accuracy comparison of decap budgets between our proposed approach (Hermite PC) and MC by using CG and SLP. The benchmarks of *mean*, *lower bound*, and *upper bound* are obtained based on MC sampling from 4000 trials. The notation, *rn*, indicates the number of random variables used in considering stochastic current sources. The *mean* is the error percentage of deterministic decap budgets against the *mean* calculated from the samples of MC. The upper or lower bound indicates the *mean* plus or minus three *standard deviations* of the node voltages. The error percentages of the decap budgets between our proposed methods against the results from MC are shown in the table. Undoubtedly, our proposed approach provides sufficient accuracy. Based on 4000 trials, the errors of percentage in comparison to MC are in the range of 0.01% and 1% in solving decap budgets with CG and SLP optimization respectively.

In consideration of the scalability of the proposed approach, Table 2 shows the speedup of the Hermite PC method over Monte Carlo method with 10,000 trials. In this table, #*node* denotes the number of nodes in the power grid circuits. *Speedup_CG* and *Speedup_SLP* represent the ratio of speedup in terms of CPU times used in CG and SLP against MC respectively. It can be seen that the proposed

Table 1. Error comparison between Hermite PC against Monte Carlo methods using CG and SLP method

	Lower bound (%)		Mean (%)		Upper bound (%)	
	1 rn	7 rn	1 rn	7 rn	1 rn	7 rn
CG	0.014	0.007	0.009	0.005	0.018	0.004
SLP	0.89	0.94	0.04	0.03	0.48	0.78

Table 2. CPU time comparison between Hermite PC and Monte Carlo

Ckt	#node	Speedup_CG	Speedup_SLP
gridrc_5	185	8152.25	7462.8
gridrc_12	3024	991.8	2718.7
gridrc_15	6105	154.8	470.03

method is about 100X to 1,000X faster than the MC method based on 10,000 trials of MC depending upon the size of the circuit and the selection of optimization algorithms.

There are two important observations in this experiment. As mentioned in Section 2, we need at least $10,000$ trials to guarantee 1% of accuracy [16]. Thus, in order to get 0.1% or even 0.01% of accuracy, we may need to have at least $1,000,000$ or even $100,000,000$ of trials. However, this is not practical in our experiment. Hence, the accuracy comparison in this experiment may not guarantee the true comparison. Nevertheless, it does show some confidence level of comparison. The other observation is that in reality the variations may not be so small as defined in our experiment. Typically the variability may be as high as 20% - 30% or even 60% of nominal DC value [1]. This means that the variability can be as high as 180 μA based on DC nominal value of 300 μA. In order to obtain an accurate result with such a huge variation, we are looking at a huge size of sampling (in the range of several millions to tens of millions) needed in order to plot a good decap optimization distribution. Thus, in order to illustrate the concept of our proposed algorithm, we limit our variations to smaller numbers with smaller size of samplings, so that the experiment can be finished in a reasonable time frame. As a result, the budgets in terms of farads may seem to be small in our experiments. However, in reality, the saving of budgets may be significant given the fact that the variability may be huge in manufacturing.

5 Conclusions

In this paper, we have proposed a fast statistical decap optimization algorithm for noise reduction of on-chip VLSI power delivery systems. The new approach is based on statistical spectral analysis method to compute the statistical voltage drops and allocate the decaps such that the resulting circuits have much higher probability to meet the voltage *IR* drop or equivalently *ground bounce* constraints statistically. The new algorithm exploits the relative constant variations for different decap configurations of on-chip power grids to speed up the statistical optimization process. The new method also considers the spatial correlations of random variables to model inter-die and intra-die variations. Experimental results show that the new approach is 100X to 1,000X faster than the brute-force Monte Carlo based statistical decap optimization. Future research may include the variations caused by temperature and/or oxide thickness during the manufacturing process.

References

[1] International technology roadmap for semiconductors(itrs), 2006 update, 2006. http://public.itrs.net.

[2] H. B. Bakoglu. *Circuits, Interconnections, and Packaging for VLSI*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1990.

[3] S. Bobba, T. Thorp, K. Aingaran, and D. Liu. IC power distribution challeges. In *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, pages 643–650, 2001.

[4] H. H. Chen and D. D. Ling. Power supply noise analysis methodology for deep-submicron VLSI chip design. In *Proc. Design Automation Conf. (DAC)*, pages 638–643, 1997.

[5] R. Downing, P. Gebler, , and G. Katopis. Decoupling capacitor effects on switching noise. *IEEE Trans. on Components, Hybrids, and Manufacturing Technology*, 16(5):484–489, Aug. 1993.

[6] J. Fan, I. Liao, S. X.-D. Tan, Y. Cai, and X. Hong. Localized on-chip power delivery network optimization via sequence of linear programming. In *Proc. Int. Symposium. on Quality Electronic Design (ISQED)*, pages 272–277, 2006.

[7] R. Ghanem. The nonlinear Gaussian spectrum of log-normal stochastic processes and variables. *Journal of Applied Mechanics*, 66:964–973, December 1999.

[8] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Dover Publications, 2003.

[9] P. Ghanta, S. Vrudhula, R. Panda, and J. Wang. Stochastic power grid analysis considering process variations. In *Proc. European Design and Test Conf. (DATE)*, volume 2, pages 964–969, 2005.

[10] K. Kang, B. C. Paul, and K. Roy. Statistical timing analysis using levelized covariance propagation. In *Proc. European Design and Test Conf. (DATE)*, 2005.

[11] T. Karnik, S. Borkar, and V. De. Sub-90nm technologies-challenges and opportunities for CAD. In *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, pages 203–206, San Jose, CA, Nov 2002.

[12] H. Li, Z. Qi, S. X.-D. Tan, L. Wu, Y. Cai, and X. Hong. Partitioning-based approach to fast on-chip decap budgeting and minimization. In *Proc. Design Automation Conf. (DAC)*, pages 170–175, June 2005.

[13] N. Mi, J. Fan, and S. X.-D. Tan. Analysis of power grid networks considering lognormal leakage current variations with spatial correlation. In *Proc. IEEE Int. Conf. on Computer Design (ICCD)*, pages 56–62, 2006.

[14] C. Paul. Effectiveness of multiple decoupling capacitors. *IEEE Trans. on Electromagnetic Compatibility*, 34(2):130–133, May 1992.

[15] C. K. S. Zhao, K. Roy. Decoupling capacitance allocation and its application to power-supply noise-aware floorplanning. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 21(1):81–92, Jan. 2002.

[16] L. Scheffer. The count of monte carlo. In *ACM/IEEE International Workshop on Timing Issues (TAU)*, 2004. http://www.tauworkshop.com.

[17] H. Su, S. S. Sapatnekar, and S. R. Nassif. Optimal decoupling capacitor sizing and placement for standard cell layout designs. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 22(4):428–436, April 2003.