A New Statistical Optimization Algorithm for Gate Sizing

Murari Mani and Michael Orshansky
Department of Electrical and Computer Engineering
University of Texas at Austin
mani@ece.utexas.edu

Abstract— **In this paper, we approach the gate sizing problem in VLSI circuits in the context of increasing variability of process and circuit parameters as technology scales into the nanometer regime. We present a statistical sizing approach that takes into account randomness in gate delays by formulating a robust linear program that can be solved efficiently. We demonstrate the efficiency and computational tractability of the proposed algorithm on the various ISCAS'85 benchmark circuits. Across the benchmarks, compared to the deterministic approach, the power savings range from $23-30\%$ for the same timing target and the yield level, the average power saving being 28%. The runtime is reasonable, ranging from a few seconds to around 10 mins, and grows linearly.**

I. Introduction

With the aggressive scaling of VLSI technology in the past decade, process variations are starting to play a significant role in the design and optimization of deep-submicron circuits [1] and can no longer be ignored. This trend can be attributed to several important factors. First, process control in the manufacturing phase is not improving at a rate comparable to scaling, causing the variability in physical dimensions, such as the effective gate length of the transistor, to proportionally increase. New systematic variation-generating mechanisms have appeared, such as the spatial channel length variation due to proximity effects and lens aberrations [2]. Second, there is an emergence of fundamental atomic-scale randomness, such as the variation in the number of dopants in the transistor channel [3]. As a result, identically designed chips are exhibiting a large spread in performance metrics severely impacting parametric yield.

The growing variability of process and circuit parameters calls for the introduction of a new statistical (stochastic) design paradigm. Much attention has been recently given to statistical timing analysis methods, specifically, statistical STA [4][5]. Optimization techniques used in the computer-aided design of VLSI circuits also need to be revamped. Indeed, the existing "deterministic" techniques rely on selecting a single number to represent the properties of a circuit under the specific conditions, for example, the gate delay for a fixed size. The options are to set the values either to the mean or the worst-case values of the varying parameters. While setting the parameters to their worst case values ensures high yield, it leads to over-conservatism in design (higher total area and power). Setting these parameters to their nominal values, however, produces unacceptable timing yield. It is clear that new methodologies are needed to tackle the problem of variability efficiently to ensure a high yield while remaining within the specified design budget for area and power. Deterministic optimization schemes, by definition, lack the explicit notion of parametric yield, preventing design for yield as an active design strategy.

Gate sizing is one of the most potent optimization techniques used in automated VLSI design. It involves making the near-critical paths faster by sizing up the gates on these paths, and sizing down gates on non-critical paths while satisfying certain constraints. The sizing problem has been formulated in several ways including unconstrained delay minimization [6], and area and power minimization under delay constraints [7]. Multiple solution methods have also been explored: TILOS [8] uses a posynomial delay model and a sensitivity-based optimization. A linear programming algorithm was used in [9], and an approach based on Lagrangian relaxation was used in [10]. Among other approaches are those based on genetic and polytope algorithms [11]. However, none of these approaches take variability into account, treating gate delays as fixed quantities. While this might have been justified in earlier technologies, it is no longer sufficient for design of high-yielding and high-performance deep-submicron circuits where the impact of variability is high.

Several previous attempts to introduce statistical considerations into circuit sizing are known to the authors [12][13]. In [12], a non-linear gate sizing problem based on a statistical gate delay model is proposed. The approach requires a number of complicated and computationally expensive maximum operations to be performed iteratively at each node. The major shortcoming of this approach is its prohibitive runtime resulting from a need to evaluate first and second order derivatives of functions that capture the dependence of delay mean and variance. An approach based on the concepts from utility theory [13] identifies the paths that are most likely to violate the timing requirements, considering both the mean delay and the variance, and tries to limit the yield loss due to these paths by sizing them appropriately. The approach is path-based, and requires a lot of pre-processing to extract the top few paths to work on, making it infeasible for large circuits. We are also aware of other current efforts to formulate a statistical circuit sizing problem [19].

In this paper, we present a new approach to statistical gate sizing. This is the first approach that is based on a completely rigorous formulation and derived from the general

principles of stochastic optimization. The original formulation is cast into a robust linear program, which is then reformulated as a Second Order Conic Program to analytically capture the dependence of the objective function on the variance of gate delays in closed form. The structure of this program allows us to achieve significantly better run-time compared to the known approaches. The algorithm has been validated on several ISCAS'85 circuits, and the results indicate that large (up to 30%) power savings are possible at the same frequency target.

The rest of the paper is organized as follows. Section II presents our model formulation. We analyze the results obtained by applying our approach to various ISCAS'85 benchmark circuits in Section III. Section IV presents the conclusions.

II. Statistical Gate Sizing: Models and Optimization

In this section we describe our formulation of the statistical gate sizing problem as it is one of the most widely used VLSI circuit optimization techniques. A variety of sizing problem formulations have appeared over the years. We consider the use of sizing for dynamic power minimization. We rely on a straightforward relationship between the total dynamic (switching) power consumption of a circuit and its total gate capacitance, and, thus, area: $P_{dyn} = CV^2f$ and $C \sim C_{ox}LW_{tot}$ where C is the total switching capacitance, C_{ox} is the oxide capacitance, f is the frequency of operation and V is the supply voltage. W_{tot} is the sum of the widths of all the transistors in the circuit, which, corresponds to the sum of sizes of the gates in the circuit. Though power and area scale by the same factor, this ignores the fact that the switching factor changes with change in circuit timing due to sizing [12].

This allows us to use a simple formulation for the power minimization problem in which the objective of gate sizing is to find an assignment of gates sizes to minimize total gate area while ensuring that the delay of the critical path is lesser than a pre-specified constraint:

$$\begin{aligned} & min \sum_i s_i \\ s.t. \quad & T_{max} \leq T \end{aligned} \tag{1}$$

where, s_i is the size of gate i, T is the specified timing target and T_{max} is the delay of the critical path through the circuit. This is the traditional gate sizing formulation, however, in our approach, area minimization is a proxy for switching power minimization.

A. Statistical Gate Sizing

Using the deterministic formulation of the sizing problem above, we can pose the sizing problem under uncertainty as:

$$\begin{aligned} & min \sum_i s_i \\ s.t. \quad & P(T_{max} \leq T) \geq \eta \end{aligned} \tag{2}$$

$$\begin{aligned} & T_{max} = max(T_o) \, \forall \, o \in PO \\ & l \leq s_i \leq u \, \forall \, i \in gates \end{aligned} \tag{3}$$

where, l and u are lower and upper bounds respectively on the size of the gates, T_o is the arrival time at output o and T is the timing target. This is a chance constrained linear program. This kind of formulation has roots in stochastic programming in which the constraint only has to be met with probability of η. In the context of circuit sizing, the parameter η corresponds to the timing yield of the circuit sized to meet the timing requirement T. It has been shown [16] that this problem is convex under the assumption of normality and for $\eta > 0.5$, and hence has a unique global minimum.

We begin the derivation by considering a single path. For a path $p \in P$, where P is the set of all paths in the circuit, delay along the path is given by:

$$d_p = \sum_i d_i \, \forall \, i \in path \, p \tag{4}$$

For random normal gate delays, we can re-write the constraint of (2) as:

$$p\left(\frac{d_p - \overline{d_p}}{\sigma_p} \leq \frac{T - \overline{d_p}}{\sigma_p} \right) \geq \eta \tag{5}$$

Recognizing that $\frac{d_p - \overline{d_p}}{\sigma_p}$ is a zero mean unit variance standard normal variable, (5) can be written as:

$$\frac{T - \overline{d_p}}{\sigma_p} \geq \phi^{-1}(\eta) \tag{6}$$

or, equivalently as:

$$\overline{d_p} + \phi^{-1}(\eta) \, \sigma_p \leq T \tag{7}$$

where $\overline{d_p}$ is the nominal delay of the path p, σ_p is the standard deviation of the delay of the path, and ϕ is the *cdf* of $N(0,1)$ [14].

What makes the statistical sizing problem qualitatively different from the deterministic one is that not only the mean path delay value but also the variance of path delay is a function of decision variables of the optimization problem (gate sizes). Capturing this dependence while keeping the optimization problem computationally tractable is the challenge of statistical gate sizing.

B. Gate Delay Modeling

An approach taken in this work is to formulate the problem as a robust linear programming problem. This is because a well-developed theory is available for solving such problems. The price that has to be paid is a linearized model of gate delay. Traditionally, a posynomial gate delay model, for example, the Elmore delay model, has been used to describe the gate delay dependence on the size of the gate [8][17]. The posynomial model describes the delay of a gate i as:

$$d_i = d_{int_i} + c\frac{\sum_j s_j}{s_i} \; j \in fanout(i) \tag{8}$$

where d_{int_i} is the intrinsic delay of gate i when it is not driving any load and c is a constant. In order to exploit

the advantages of linear optimization techniques, such as the ability to find a global optimum and the existence of fast optimization algorithms (e.g. simplex), a linear gate delay model has been proposed in [9]. We adopt such a model in this work. The gate delay is given by:

$$d_i = a_i - b_i s_i + c_i \sum_{j \in fo(i)} s_j \tag{9}$$

Here, (a_i, b_i, c_i) are fitting coefficients that can be empirically determined via SPICE-based circuit simulation for each gate in the library.

C. Computationally Efficient Formulation

In translating the problem from a conceptual problem to a computationally tractable one, we need to address the two concerns of explicitly and analytically encoding the dependence of variance on the decision variables, and translating a path-based formulation into a node-based formulation.

Using the linear delay model defined above and (4), the deterministic sizing problem of (1) can be rewritten as a linear programming (LP) problem:

$$min \sum_i s_i$$

$$d_p = \sum_{i \in p} \left(a_i - b_i s_i + c_i \sum_{j \in fo(i)} s_j \right)$$

$$s.t. \qquad d_p \leq T \, \forall \, p \in P \tag{10}$$

We now pose a statistical (robust) counter-part of the above LP, which allows us to bring uncertainty into the problem. We define the coefficients of the linear delay model to be uncertain (random) variables. Specifically, we model the variability in gate delay by assuming that b_i, c_i are normal random variables. Then, the robust LP is:

$$min \sum_i s_i$$

$$d_p = \sum_{i \in p} \left(a_i - b_i s_i + c_i \sum_{j \in fo(i)} s_j \right)$$

$$s.t. \qquad p(d_p \leq T) \geq \eta \, \forall \, p \in P \tag{11}$$

In order to make the formulation above computationally efficient, under the assumption of node delay independence and for node delays which are Gaussian random variables, we can re-express the probabilistic constraint as (12). This specific formulation can then be efficiently solved [14][15]. While the assumption of independence appears to be essential to this problem formulation, the node distributions do not have to be Gaussian. Any other arbitrary distribution can be used by appropriately selecting the inverse *cdf* function $\phi^{-1}(\eta)$.

$$min \sum_i s_i$$

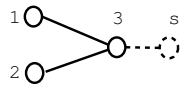

Fig. 1. Example to illustrate statistical sizing algorithm

$$d_p = \sum_{i \in p} \left(a_i - b_i s_i + c_i \sum_{j \in fo(i)} s_j + \phi^{-1}(\eta) \sqrt{\sigma_{b_i}^2 s_i^2 + \sigma_{c_i}^2 (\sum_{j \in fo(i)} s_j)^2} \right)$$

$$s.t. \qquad d_p \leq T \, \forall \, p \in P \tag{12}$$

where σ_{b_i} and σ_{c_i} are the variances of b_i and c_i respectively.

Still another problem to address is translating a path-based formulation into a node-based formulation for large circuits, with the number of paths growing exponentially. As it is not feasible to enumerate them, we transform the path based constraints into node-based constraints. The delay of a gate is now given by:

$$\hat{d}_i = \left(a_i - b_i s_i + c_i \sum_{j \in fo(i)} s_j + \phi^{-1}(\eta) \sqrt{\sigma_{b_i}^2 s_i^2 + \sigma_{c_i}^2 (\sum_{j \in fo(i)} s_j)^2} \right) \tag{13}$$

We propagate $\hat{d}_i$ through the circuit using the conventional arrival time equations of static timing analysis. This is illustrated using the simple circuit shown in Fig. 1. Nodes 1 and 2 are the primary inputs and 3 is the primary output. Node s is the dummy sink node. The node-based statistical sizing formulation for this circuit is:

$$min \sum_{i=1}^{3} s_i$$

$$\begin{aligned} s.t. \qquad & \hat{d}_1 \leq t_3 \\ & \hat{d}_2 \leq t_3 \\ & t_3 + \hat{d}_3 \leq T \end{aligned} \tag{14}$$

where, t_3 is the arrival time at node 3 and the $\hat{d}_i$s are given by (13). Though this transformation introduces some sub-optimality, we found it to be of the order of $1 - 2\%$ for smaller circuits, consisting of up to 20 levels of logic. Sub-optimality here refers to the decrease in dynamic power achieved by the path-based formulation compared to the node-based implementation for the same timing constraint and yield level. In general, a simple analysis shows that the sub-optimality increases as $O(l)$, where l is the logic depth of the circuit. However, given the trend towards shallower circuits, we believe that the overall impact is not going to be significant. In any case, we believe that the computational benefits in terms of run-time justify such a transformation.

TABLE I

MINIMUM POWER OBTAINED BY DETERMINISTIC AND STATISTICAL ALGORITHMS AT DIFFERENT YIELD LEVELS

Circuit	No. of gates	T_{target}	Det. sizing	Stat. sizing			VSS
			P_{det}	$P_{99.7\%}$	$P_{96.4\%}$	$P_{84\%}$	
C432	160	1034.4	348.1	266.53	202.54	177.63	23.5
C499	202	850.9	544.1	391.34	309.2	268.2	28.1
C880	383	1002.39	740.3	490.3	456.8	438.4	33.7
C1355	880	1049.7	1575.0	1099.5	822.2	730.3	30.22
C1908	1193	1466.4	1460	1137.59	1052.98	1004.45	22.1

Another important issue is the selection of margin coefficients (the value of $\phi^{-1}(\eta)$) for each of the gates. We adapt a simple approach and set the margin coefficients to $\phi^{-1}(\eta)$ for each node, where η is the required yield. This is an *ad hoc* procedure, but it seems to approximate the actual yield of the circuit quite well (Fig. 2). However, it is possible to adopt a much more sophisticated line search technique, at the expense of computational complexity. This will involve iterating over the solution space and progressively refining the values of the coefficients subject to the yield obtained from the particular sized configuration [19].

III. RESULTS AND ANALYSIS

We implemented our optimization using the General Algebraic Modeling Software (GAMS) [18]. The statistical sizing model was solved using the CPLEX Successive Linear Program (CPLEXSLP) solver and the deterministic model using the CPLEX LP solver. Both the deterministic and statistical optimization problems are convex and hence we are assured of a globally optimal solution. We obtained delay equation coefficients for all the gates that appear in the ISCAS85 benchmark circuits by characterizing the delay values for the various gate sizes and fanouts for a $0.13\mu m$ CMOS process. The overall variation in gate delay is assumed to be 25% (in the sense of 3σ/mean). This is translated into the specific variability of b and c. Though (a_i, b_i, c_i) are fitting coefficients and the actual physical variability they map to is not clear, we choose a lower level of variation for c_i because we feel that the coefficient b_i maps to a larger underlying physical variability (largely corresponding to the variation in the channel length, L_{eff}) than c_i (largely corresponding to the variation in oxide thickness). We assume a standard deviation of 8% and 5% of nominal values respectively for b_i and c_i. The empirical fitting of the coefficients is done using SysStat [20]. To ensure accuracy within a working range, we restrict the sizes to $1x$ to $4x$ of the minumum size. We believe that the constraint imposed on the gate size range is reasonable, and would not lead to significant sub-optimality of the solution. The accuracy of the fit can be improved by adopting a piecewise linear function similar to [9]. In our case, within this restricted range of gate size values, the linear model is reasonably accurate, the rms error of the fit being $5 - 7\%$.

We validate the performance and run-time behavior of our optimization algorithm on several of the ISCAS'85 benchmark circuits. (Because of the memory limitations on the servers that we ran our optimizations on, we weren't able to apply our algorithm to the larger benchmark circuits. However, this is purely a logistical issue that can be resolved and is not a fundamental limitation to our algorithm with regards to the run-time or the optimality of the solution). The power savings that our approach can enable without the loss of performance or yield are documented in Table I. The way to interpret the results is as follows. We perform a linear optimization for the circuits where the random parameters are set to their worst case values. T_{target} corresponds to the minimum delay through the circuit obtained by unconstrained optimization in this deterministic setting and P_{det} is the corresponding power. $P_{99.7\%}$ is the power from the statistical sizing algorithm when the timing constraint is T_{target} for the 99.7% yield level. The value of stochastic solution (VSS) is defined as

$$VSS = \frac{P_{det} - P_{99.7\%}}{P_{det}} 100. \tag{15}$$

As can be seen, we obtain a sizable saving in power by applying our approach.

Table II presents a comparison of execution times for our optimization scheme with those for the deterministic scheme. For all circuits, the run time is on the order of a few minutes or lesser. Fig. 3 serves to demonstrate the fact that the run-time is roughly linear in the number of nodes. Comparing our run-time with those reported in [12], for circuits with comparable number of nodes, the run-time of our statistical sizing algorithm is an order of magnitude better.

As mentioned in Section II, we used a simple approach to achieve the varying levels of parametric yield: the same value of the margin coefficient was used for all the nodes. In order to validate this strategy we carried out the following experiment. Fig. 2 shows the result of performing a Monte - Carlo analysis on the nominal configuration. 1000 simulations were used. It depicts that the actual yield very closely approximates the yield predicted by our optimization. This justifies choosing the same value of the margin coefficients for all the nodes in the circuit.

Figures 4-6 show a set of Pareto curves for the $C432$ benchmark. Fig. 4 plots the objective function (power) vs. the required arrival time at the output for various timing yield levels. In the represented power-delay space, the difference in power between the circuits sized at different yield levels is much greater for tighter timing constraints. Specifically, when a deterministic sizing is performed by setting the parameters to their worst case values, the sub-optimality is very large for

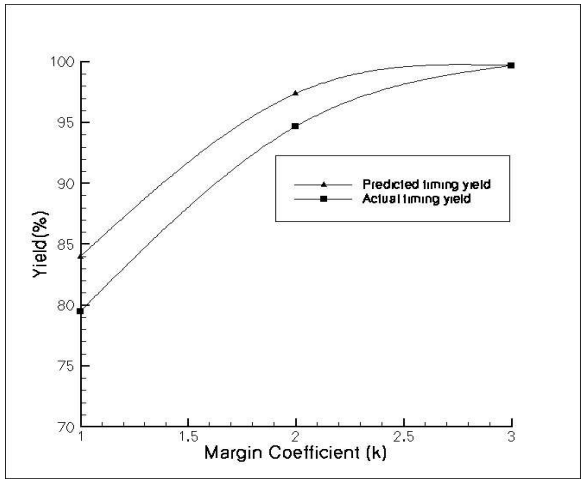

Fig. 2. The predicted yield reasonably well approximates the actual yield.

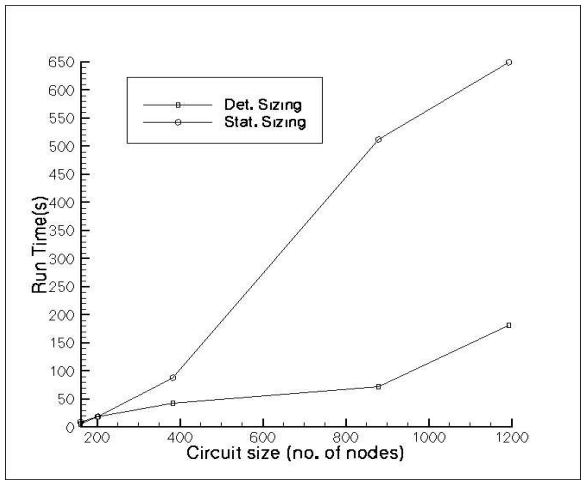

Fig. 3. A Comparison between the run-time of statistical and deterministic sizing algorithms. Both grow approximately linearly.

TABLE II
COMPARISON OF EXECUTION TIMES

Circuit	Deterministic Sizing	Statistical Sizing
C432	$5.72s$	$9.33s$
C499	$18.7s$	$19.1s$
C880	$42.5s$	$1m29s$
C1355	$1m13s$	$8m32s$
C1908	$3m02s$	$10m50s$

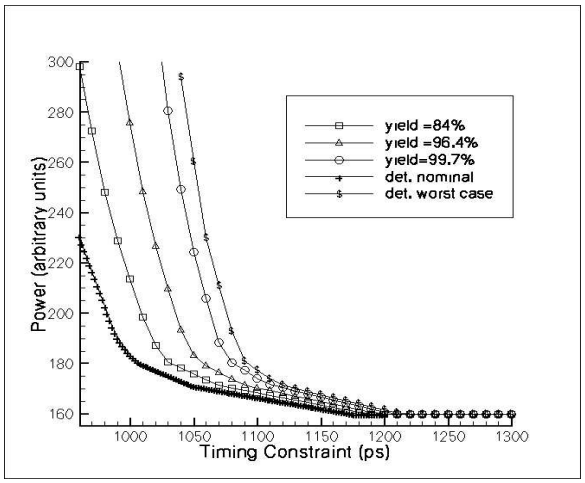

Fig. 4. The power-delay Pareto curves at different yield levels. Statistical optimization does uniformly better than the deterministic optimization at the same yield level.

tight timing constraints. However, the overhead in terms of power is low for lax timing constraints. Fig. 4 also points to the fact that in the presence of variability, certain timing targets are unachievable for a particular yield level, and designing for the nominal values of the varying parameters will lead to an unacceptably low yield. Again, this penalty grows as we approach the maximum frequency of operation of the circuit. Variability also shifts this maximum frequency towards smaller values.

Fig. 5 shows the plot of circuit yield and power for the $C432$ benchmark circuit. As expected, for relaxed timing constraints the circuit power is very insensitive to the yield requirement. However, the curve gets steeper for tighter timing requirements and the power overhead, to hit a certain yield target, grows drastically.

To study how the magnitude of variability affects the relationships between yield, timing and circuit power, we changed the standard deviation (σ) of the parameter values from 5% to 12% of the nominal value. Fig. 6 shows the results of plotting

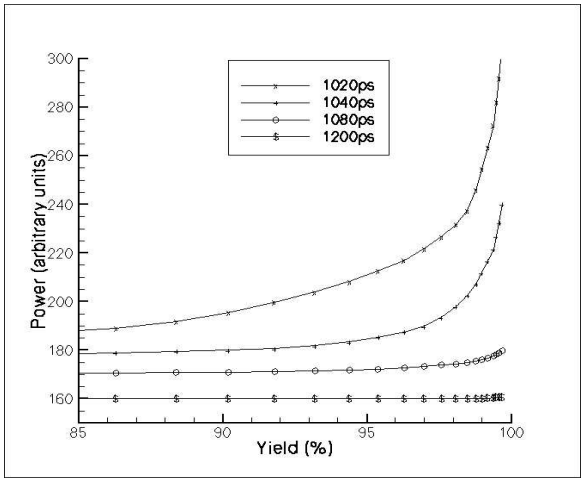

Fig. 5. The sensitivity of power to yield level. The power-delay trade-off becomes extremely unfavorable at tighter T_{req} and higher yield levels.

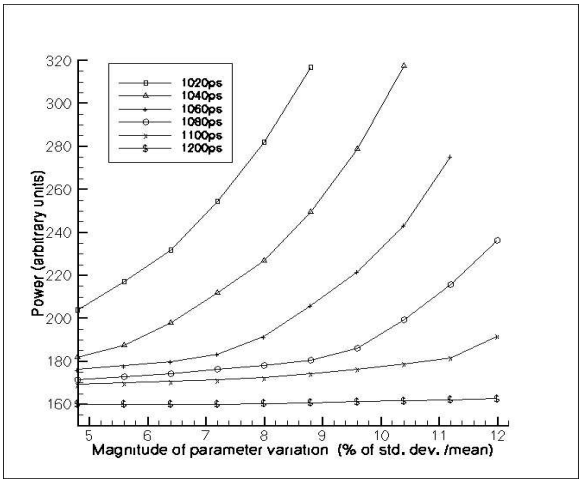

Fig. 6. The minimum achievable power goes up for higher magnitude of parameter variation ($\sigma/mean$), especially, for tight timing constraints.

the magnitude of standard deviation against circuit power for different required times. This graph underscores the fact that the impact of process variations is strongly dependent on the timing specification required of the circuit and the overhead is most significant at tighter timing constraints.

IV. Conclusion

In this paper we have taken a step in the direction of true statistical gate sizing. A statistical solution performs uniformly better than the non-statistical solutions in terms of the achievable power and delay values. We show that significant power savings (up to 30%) are possible by applying our approach. The runtime of our algorithm is significantly better than that of the known alternative schemes.

Acknowledgment

We would like to thank Prof. Morton of UT Austin's Department of Operations Research for helpful discussions.

References

[1] C. Visweswariah, "Death, taxes and failing chips", *Proc. of Design Automation Conference*, 2003, pp. 343-347.
[2] M. Orshansky, L. Milor, C. Hu, "Characterization of Spatial Intrafield Gate CD Variability, Its Impact on Circuit Performance, and Spatial Mask-Level Correction", *IEEE Transactions on Semiconductor Manufacturing*, Volume:17, Issue:1, Feb. 2004, pp. 2-11.
[3] Y. Taur et al, "CMOS scaling into the nanometer regime", *Proceedings of the IEEE*, Volume:85, Issue:4, April 1997, pp. 486-504.
[4] M. Orshansky and K. Keutzer, "A general probabilistic framework for worst case timing analysis", *Proc. of Design Automation Conference*, June 2002, pp. 556-561.
[5] A. Devgan and C. Kashyap, "Block-based static timing analysis with uncertainty", *Proc. of Intl' Conf. on CAD*, 2003, 9-13, pp. 604-614.
[6] O.Coudert, "Gate sizing: A general purpose optimization approach", *Proc. of EDTC'96*, Paris, France, 1996.
[7] M. Borah, R.M. Owens, M.J. Irwin, "Transistor sizing for minimizing power consumption of CMOS circuits under delay constraint", *Proc. of Int'l symposium on low power design*, Dana Point, April 1995, pp. 167-172.
[8] J. P. Fishburn and A. E. Dunlop, "TILOS: A posynomial programming approach to transistor sizing", *IEEE Trans. on CAD*, pp. 326-336.
[9] M. Berkelaar, J. Jess, "Gate sizing in MOS digital circuits with linear programming", *Proc of the European Design Automation Conference*, 1990, 12-15 March 1990, pp. 217-221.
[10] C.P. Chen, C.C.N Chu and D.F. Wong, "Fast and exact simultaneous gate and wire sizing by Lagrangian relaxation", *IEEE Trans. on CAD*, 1999, pp. 1014-1025.
[11] O. Coudert, R. Haddad, S. Manne, "New algorithms for gate sizing: a comparative study", *Proc. of Design Automation Conference*, 1996, 3-7 June 1996, pp. 734-739.
[12] E. T. A. F. Jacobs and M. R. C. M. Berkelaar, "Gate sizing using a statistical delay model", *Proc. of IEEE/ACM Design Automation and Test Conf.*,2000, pp. 283-290.
[13] S. Raj, S. Vrudhula, J. Wang, "Statistical Gate Sizing to increase Timing Yield", *Proc. of TAU'04*, Austin, TX, 2004.
[14] S. Boyd, L. Vandenberghe, *Convex Optimization*.
[15] A. Prekopa, *Stochastic Programming*, Kluwer Academic, 1995.
[16] P. Kall, S. Wallace, *Stochastic Programming*, www.unizh.ch/ior/Pages/Deutsch/Mitglieder/Kall/bib/ka-wal-94.pdf.
[17] S. S. Sapatnekar, V. B. Rao, P. M. Vaidya, and S. M. Kang, "An exact solution to the transistor sizing problem for CMOS circuits using convex optimization", *IEEE Transactions on Computer-Aided Design*, vol. 12, Nov. 1993, pp. 1621-1634.
[18] http://www.gams.com/docs/gams/GAMSUsersGuide.pdf
[19] S. Boyd, private communication, March 2004.
[20] http://www.systat.com/products/TableCurve