Leakage Minimization of Nano-Scale Circuits in the Presence of Systematic and Random Variations* †

Sarvesh Bhardwaj
Computer Science and Engineering
Arizona State University
sarvesh.bhardwaj@asu.edu

Sarma B.K.Vrudhula
Computer Science and Engineering
Arizona State University
sarma.vrudhula@asu.edu

ABSTRACT

This paper presents a novel gate sizing methodology to minimize the leakage power in the presence of process variations. The leakage and delay are modeled as posynomials functions to formulate a geometric programming problem. The existing statistical leakage model of [18] is extended to include the variations in gate sizes as well as systematic variations. We propose techniques to efficiently evaluate constraints on the α-percentile of the path delays without enumerating the paths in the circuit. The complexity of evaluating the objective function is $O(|N|^2)$ and that of evaluating the delay constraints is $O(|N| + |E|)$ for a circuit with $|N|$ gates and $|E|$ wires. The optimization problem is then solved using a convex optimization algorithm that gives an *exact* solution.

Categories and Subject Descriptors

B.6.3 [**Logic Design**]: Design Aids—*Optimization*

General Terms

Algorithms, Design, Performance

Keywords

Leakage, Statistical, Optimization, Geometric Programming

1. INTRODUCTION

The lack of process uniformity in the semiconductor manufacturing has caused variability to become the primary cause of concern for nanometer scale CMOS design. Significant research efforts have focused on understanding the causes and effects of spatial variations [24, 23, 13, 1, 10, 15, 26]. The variations are caused by either global effects [24] such as mask imperfections and lens aberration, or local effects [9] such as layout pattern variations.

As shown in [1], for 30% variations in the circuit delay there can be 20X variations in the leakage current. Various design techniques for leakage reduction such as transistor stacking [12], sleep transistor insertion [8], body biasing [14] and driving the circuit into a minimum leakage sleep state have been proposed in the past. The power savings accrued by these techniques can be supplemented by gate sizing and dual-threshold voltage (V_t) assignment [29, 20, 7, 21].

Traditional deterministic gate sizing [19, 3] and dual-V_t assignment techniques [6] can be classified into two major categories: discrete optimization approaches and non-linear optimization techniques. In discrete optimization techniques, starting with an initial feasible implementation (say, all gates assigned the smallest gate size and low-V_t), the gates to be sized up (or assigned a higher V_t) are selected based on their sensitivities. In every iteration, the timing constraints are checked for any possible violations. These techniques provide good solutions but because of large number of feasible solutions, it is difficult to guarantee the optimal solution. Non-linear programming based techniques use suitable models for the objective function (power) and the constraints (delay) to formulate an optimization problem. This problem can then be solved using non-linear optimization techniques [5, 19] to obtain an optimal solution.

In the presence of process variations, the parameters have to be modeled as random variables. Due to this, there has been an increased interest in techniques for statistical analysis [27] and optimization of circuit performance. A number of statistical optimization methods ([4, 25] to name a few) have been proposed in the past. These approaches are primarily based on evaluating the yield (probability that a design has acceptable performance) as the integral of the joint probability distribution (jpdf) of the circuit performance ϕ over its *acceptability region*, A_ϕ. Once an approximation to A_ϕ is obtained, the design center is moved to the interior of A_ϕ such that the yield is maximized. Because of their high computational complexity, these techniques are suitable only for small circuits where the number of design parameters is small. Another class of statistical optimization techniques [22, 17] have been proposed recently. These techniques are similar to the deterministic optimization techniques described above but they model the variability in the device parameters.

In this paper we present a statistical optimization approach based on modeling the statistics (mean and second moment) of leakage and delay as *posynomial functions* [16] of nominal gate sizes. A function of mean and variance of leakage is minimized subject to constraints on α-percentile of the delay. The posynomial functions can be transformed into convex functions by a variable transformation. This

*Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work was carried out at the NSF's State/Industry/University Cooperative Research Centers' (NSF-S/IUCRC) Center for Low Power Electronics (CLPE). CLPE is supported by the NSF (Grant #EEC-9523338), the State of Arizona, and an industrial consortium. This work was also supported by NSF through grant #CCR-0205227.

†Full version of this paper is available at:http://veda.eas.asu.edu/papers/bhardwaj-dac05.pdf

DAC 2005, June 13–17, 2005, Anaheim, California, USA.

convex optimization problem is then solved to obtain a globally optimal solution [19].

The organization of rest of the paper is as follows: The problem is formally defined in Section 2. Section 3 describes the statistical leakage and delay models used in the formulation. The proposed solutions is discussed in detail in Section 4. Finally the experimental results and conclusions are described in Sections 5 and 6.

2. PROBLEM FORMULATION

Let a circuit be represented using a Directed Acyclic Graph (DAG) $G = (N, E)$, where $N = \{1, 2, .., n\}$ is the set of nodes and $E = \{(i, j) : i, j \in N\}$ is the set of edges. The nodes correspond to the gates in the original circuit. An edge (i, j) represents that gate i fanouts to gate j.

Let the parameter space for each gate i be defined as $\hat{u}_i = (u_1^i, u_2^i, .., u_r^i)$, where r denotes the number of parameters. In the presence of process variations, each of these parameters is a random variable. Hence, if Ω denotes the space of manufacturing outcomes, $\hat{u}_i : \Omega \rightarrow \mathbb{R}^r$ is a function that maps every outcome $\omega \in \Omega$ to a point in an r-dimensional Euclidean space. Hence, the parameters for the manufacturing outcome ω are given by $\hat{u}_i(\omega) = (u_1^i(\omega), u_2^i(\omega), .., u_r^i(\omega))$. The parameters considered in this work include the gate length (L_i), threshold voltage (V_i), the gate size (w_i) and the oxide thickness (T_i). For simplicity, the parameters are modeled as independent random variables. Other parameters such as supply voltage (V_{dd}) are modeled as deterministic quantities. For convenience, henceforth the explicit dependency of $\hat{u}$ on the argument ω will not be shown.

The circuit leakage and delay under this variational model are also random variables as they are functions of random parameters. Let I_T denote the total leakage of the circuit and D_p denote the delay of path $p \in \mathcal{P}$, where $\mathcal{P}$ represents the set of paths in the circuit. The stochastic leakage minimization problem can now be formulated as follows

$$\min_{\omega \in \Omega} \quad I_T(\hat{u}_1, \hat{u}_2, .., \hat{u}_n, \omega) \tag{1}$$

$$\text{sub. to} \quad \mathbf{P}(D_p(\hat{u}_1, .., \hat{u}_n, \omega) \leq T_{req}) \geq \alpha \quad \forall\, p \in \mathcal{P}. \tag{2}$$

where $\mathbf{P}(X \leq x)$ denotes the probability that the random variable X is less than or equal to x. α can be considered to be a *confidence level*. As the number of manufacturing outcomes ω can be infinite, it does not make sense to solve the optimization problem for every ω. Moreover, solving for a single ω will give the optimal choice of parameters for that particular manufacturing outcome (the probability of which is close to zero). Hence, a more relevant objective would be some statistic (such as *mean* or *variance*) of the leakage current. From Figure 1, it can be seen that minimizing the expected value of the leakage without any regard to its variance results in an increased number of chips having lower frequency (curve A). Whereas, minimizing just the variance without optimizing the mean leaves a scope of reduction in the leakage of the manufactured circuits (curve B). Hence the goal of maximizing the leakage yield can be achieved by minimizing a linear combination of the square of the mean and the variance of leakage (curve C). Thus, the new objective becomes $\lambda\,\mu^2(I_T) + (1-\lambda)\,\sigma^2(I_T)$, where $\mu(X) = \mathbf{E}[X]$ and $\sigma^2(X) = \mathbf{E}[(X - \mathbf{E}(X))^2]$ for any random variable X. $\lambda \in [0, 1]$. The constraints on the path delay probability as shown in (2) can be translated into constraints on the α-percentile ($\mathbf{z}_\alpha[D_p]$) of the path delay. $\mathbf{z}_\alpha[X]$ is defined as the smallest value of the random variable X for which its cumulative distribution function is greater than $0.01 \times \alpha$.

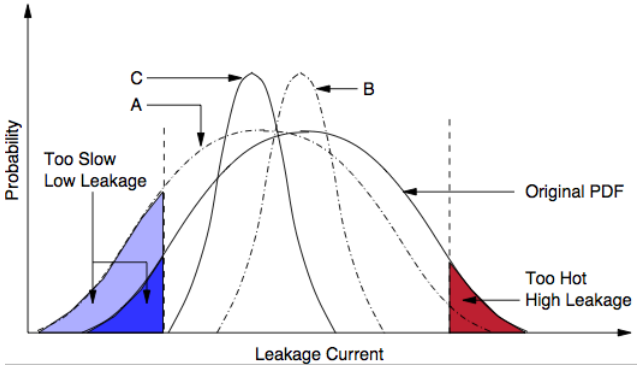


Figure 1: Leakage Reduction

The stochastic leakage minimization problem can now be written as a deterministic optimization problem in the following form

$$\min \quad \lambda\mu^2(I_T(\hat{u}_1, .., \hat{u}_n, \omega)) + (1-\lambda)\sigma^2(I_T(\hat{u}_1, .., \hat{u}_n, \omega)) \tag{3}$$

$$\text{sub to} \quad \mathbf{z}_\alpha[D_p] \leq T_{req} \qquad \forall\, p \in \mathcal{P}. \tag{4}$$

The model parameter u is modeled as $u = u_o + u_s + u_\xi$, where u_o is the value of the parameter specified by the designer (e.g. gate sizes) or determined by the technology (e.g. gate length), u_s is the systematic component of the variations and $u_\xi \sim N(0, \sigma^2(u_o))$ is the random component of the variations. $N(0, \sigma^2(u_o))$ represents a normal random variable with 0 mean and variance $\sigma^2(u_o)$. u can then be written as $u \sim N(u_o + u_s, \sigma^2(u_o))$, where $\sigma^2(u_o)$ is the variance of the parameter u as a result of specifying the value u_o. Hence u_o are the decision variables of the optimization problem. In this work, the only decision variables considered are the gate sizes ($w_{o,i}$).

There seems to be a consensus in the design as well as EDA community that the variations in a parameter increase if the nominal value of that parameter is decreased. Based on this, we model the variance of the circuit size w_i as being inversely proportional to its nominal value $w_{so,i}$. Thus

$$\sigma(w_i) = \frac{k_w}{w_{so,i}^l} \tag{5}$$

where k_w and l are model parameters.

A circuit can potentially have exponential number of paths. Section 3.2 discusses how to overcome this problem by modifying the set of constraints so that *all* the constraints can be replaced by a *single* constraint. This single constraint can be obtained by performing a single PERT-like traversal of the circuit DAG.

3. LEAKAGE AND DELAY MODELS

Recent advances in the statistical models for leakage [18] have provided an opportunity for development of statistical optimization techniques. In this work, the leakage model of [18] is extended to accommodate the variations in the gate sizes as well as include the effect of systematic variations. The objective function of the optimization problem (3) under this model is shown to be a posynomial function of gate sizes. For delay, the Elmore delay model is used.

3.1 Statistical Leakage Model

Let I_{Si} and I_{Gi} denote the sub-threshold leakage and the gate leakage of gate i respectively. The total leakage, I_{Ti} for gate i, is the sum of the sub-threshold leakage and the gate

leakage as shown in Equation 6.

$$I_{Ti} = I_{Gi} + I_{Si}. \tag{6}$$

Other sources of leakage such as Band-to-Band Tunneling (BTBT) through the reverse biased source-substrate and drain-substrate junctions will also become prominent in future technologies [11]. However, the current work only targets the sub-threshold leakage and gate leakage because statistical models have already been developed for them. The model developed in [18] assumed the sub-threshold leakage to be dependent on the gate length and threshold voltage. The gate leakage has an exponential dependence on the gate oxide thickness. Hence the total leakage of a circuit can be written as:

$$I_T = \sum_{i \in N} w_i \left(I_{So} \exp\left(\frac{-(L_i + c_2 L_i^2 + c_3 V_i)}{c_1}\right) + I_{Go} \exp\left(\frac{-T_i}{\beta}\right)\right). \tag{7}$$

where

I_{So} : Nominal sub − threshold leakage,
I_{Go} : Nominal gate leakage,
$c_1,\ c_2,\ c_3,\ \beta$: Fitting parameters.

The parameters can be expanded in terms of their systematic and random components to obtain the total leakage as shown in Equation (8).

$$I_T = \sum_{i \in N} w_i \left(I'_{So,i} \exp\left(\frac{-(\alpha_i L_{\xi,i} + c_2 L_{\xi,i}^2 + c_3 V_{\xi,i})}{c_1}\right) + I'_{Go,i} \exp\left(\frac{-T_{\xi,i}}{\beta}\right)\right). \tag{8}$$

where

$$I'_{So,i} = I_{So} \exp\left(\frac{-(L_{so,i} + c_2 (L_{so,i})^2 + c_3 V_{so,i})}{c_1}\right) \tag{9}$$

$$I'_{Go,i} = I_{Go} \exp\left(\frac{-(T_{o,i} + T_{s,i})}{\beta}\right) \tag{10}$$

$$\alpha_i = 1 + 2c_2(L_{o,i} + L_{s,i}). \tag{11}$$

The leakage currents $I'_{So,i}$ and $I'_{Go,i}$ are dependent only on the systematic component of the variations, $L_{so,i} = L_{s,i} + L_{o,i}$ and $V_{so,i} = V_{s,i} + V_{o,i}$. This work models the global variations in the gate length and the threshold voltage. Thus, the random variables corresponding to the parameters of the individual gates can be replaced by a single random variable for every parameter. Hence $L_{\xi,i} = L_\xi$, $V_{\xi,i} = V_\xi$ and $T_{\xi,i} = T_\xi$ for $i \in N$. As a result of this, both the leakage components can be written as product of two factors, one common to *all* the gates in the circuit and the other factor dependent on the individual gate size w_i.

$$I_S = \left(\sum_{i \in N} w_i I'_{So,i} \exp\left(\frac{-\alpha_i L_\xi}{c_1}\right)\right) \exp\left(\frac{-(c_2 L_\xi^2 + c_3 V_\xi)}{c_1}\right) \tag{12}$$

$$I_G = \left(\sum_{i \in N} w_i I'_{Go,i} \right) \exp\left(\frac{-T_\xi}{\beta}\right). \tag{13}$$

Thus both gate leakage and sub-threshold leakage have the form $Z = X \cdot Y$ where X and Y are independent random variables. Hence, their mean and second moment can be computed using Equations (14) and (15).

$$\mathbf{E}[Z] = \mathbf{E}[X] \cdot \mathbf{E}[Y] \tag{14}$$

$$\mathbf{E}[Z^2] = \mathbf{E}[X^2] \cdot \mathbf{E}[Y^2]. \tag{15}$$

For the random variable dependent only on w_i, the moments can be obtained as

$$\mathbf{E}\Big[\sum_{i \in N} k_i w_i\Big] = \sum_{i \in N} k_i \, \mathbf{E}[w_i] = \sum_{i \in N} k_i w_{so,i} \tag{16}$$

$$\mathbf{E}\left[\left(\sum_{i \in N} k_i w_i\right)^2\right] = \sum_{i \in N} k_i^2 \, \mathbf{E}[w_i^2] + \sum_{\substack{i,j \in N \\ i \neq j}} k_i k_j w_{so,i} w_{so,j} \tag{17}$$

where $k_i = I'_{So,i}$ for sub-threshold leakage and $k_i = I'_{Go,i}$ for the gate leakage component. Also, $w_{so,i} = w_{o,i} + w_{s,i}$ and $\mathbf{E}[w_i^2] = \sigma^2(w_i) + w_{so,i}^2$. The second component of the leakage (that is independent of the gate sizes), has the form $U = \exp(-(W + aW^2)/b)$ where $W \sim N(0, \sigma_W^2)$. Its moments can be computed using Equations (18) and (19). [18] also utilizes these equations for computing the statistics of leakage.

$$\mathbf{E}[U] = \left(1 + \frac{2a}{b}\sigma_W^2\right)^{-\frac{1}{2}} \cdot \exp\left(\frac{\sigma_W^2}{2b^2 + 4\sigma_W^2 ab}\right) \tag{18}$$

$$\mathbf{E}[U^2] = \left(1 + \frac{4a}{b}\sigma_W^2\right)^{-\frac{1}{2}} \cdot \exp\left(\frac{2\sigma_W^2}{b^2 + 4\sigma_W^2 ab}\right). \tag{19}$$

Hence the mean and variance of total leakage can be computed by using Equations (12) through (19) and performing some algebraic manipulation. The complexity of computing the mean of the leakage is $O(|N|)$ and that of computing the variance of the leakage is $O(|N|^2)$.

3.2 Statistical Delay Model

The on-resistance of a gate i is inversely proportional to its size w_i. Thus, using the Elmore delay model, the delay of gate i, d_i can be written as

$$d_i = \alpha_i + \beta_i \frac{C_{ld,i}}{w_i} \tag{20}$$

where α_i is a fitting parameter, $C_{ld,i}$ is the capacitive load at the output of gate i. The load capacitance (input capacitance of the fanout gates) is directly proportional to the sizes of the fanout gates. Hence, $C_{ld,i} = c \cdot \sum_{j \in FO(i)} w_j$, where $FO(i)$ represents the set of fanout gates for gate i. c is the capacitance per unit size. β_i captures the dependency of the delay on L_i and V_i. This dependency is modeled up to its first order approximation. That is:

$$\beta_i = \beta_{L_i} L_i + \beta_{V_i} V_i \tag{21}$$

β_{L_i} and β_{V_i} correspond to the sensitivity of the delay d_i to L_i and V_i. In the presence of variations, w_i, L_i, V_i and d_i are all random variables. Hence Equations (20) and (21) can be used to obtain the mean of the gate delay as:

$$\mathbf{E}[d_i] = \alpha_i + \mathbf{E}[\beta_i]\mathbf{E}\left[\frac{1}{w_i}\right]\left(\sum_{j \in FO(i)} c \cdot \mathbf{E}[w_j]\right) \tag{22}$$

It can be seen from Equation (22) that all the expectations can be easily computed except $\mathbf{E}[w_i^{-1}]$ as the expectation of the reciprocal of a normal random variable does not exist. We make a simplifying assumption here that $\mathbf{E}[w_i^{-1}] = \mathbf{E}[w_i]^{-1} = w_{so,i}^{-1}$. The error involved in this approximation is around 3-4%. This error is quantified in the Appendix. Thus

$$\mathbf{E}[d_i] = \alpha_i + \mathbf{E}[\beta_i] \frac{c \cdot \sum_{j \in FO(i)} w_{so,j}}{w_{so,i}} \tag{23}$$

From Equation (20) it is not possible to compute the variance of the gate delay. Hence, we use a simple model for the

standard deviation of the gate such that the standard deviation of the gate delay is directly proportional to the standard deviation of the sizes of its fanout gates (this is a reasonable assumption as the variations in the load is expected to increase the variations in the delay). Also the standard deviation is assumed to be inversely proportional to its nominal size (δ captures this effect). Hence

$$\sigma(d_i) = \gamma_i \frac{\sum_{j \in FO(i)} \sigma(w_{so,j})}{w_{so,i}^{\delta}} \tag{24}$$

These models for mean and standard deviation are used for computing the α-percentile, $\mathbf{z}_\alpha[D_p]$ of a path delay, D_p. Under the assumption of the path delays being random variables, the α-percentile of a path delay can be written as

$$\begin{aligned}\mathbf{z}_\alpha[D_p] &= \mu(D_p) + z_\alpha \sigma(D_p) \\ &= \sum_{i \in p} \mu(d_i) + z_\alpha \left(\sum_{i \in p} \sigma^2(d_i) + \sum_{\substack{i,j \\ i \neq j}} C_{ij} \right)^{\frac{1}{2}}\end{aligned} \tag{25}$$

where $C_{ij} = \rho_{ij}\sigma(d_i)\sigma(d_j)$ is the covariance of gate delays d_i and d_j. $|\rho_{ij}| \leq 1$ is the correlation coefficient of d_i and d_j. z_α is the α-percentile of the standard normal random variable ($N(0,1)$). From Equation (25), computations of the largest α-percentile in a circuit, requires the enumeration of all the paths and then computing their α-percentiles. This is a computationally expensive procedure. Hence in order to check whether a particular assignment of the gate sizes satisfies the delay constraints, we utilize Theorem 3.1.

THEOREM 3.1. *Let $w = (w_1, \ldots, w_{|N|})$ be the vector of the assigned gate sizes to the gates in a circuit. Given the feasible set of sizes, $S = \{w \in \mathbb{R}^{|N|} : \mathbf{z}_\alpha[D_p(w)] \leq T_{req},\ \forall p \in \mathcal{P}\}$ and an upper bound $\mathbf{z}_\alpha^U[D_p]$ on the α-percentile of D_p. Then $S_U \subseteq S$, where S_U is the feasible set of sizes defined by the constraints on the upper bound $\mathbf{z}_\alpha^U[D_p]$.*

PROOF. Let $w \in S_U$, this implies $\mathbf{z}_\alpha^U[D_p(w)] \leq T_{req},\ \forall p \in \mathcal{P}$. But $\mathbf{z}_\alpha[D_p(w)] \leq \mathbf{z}_\alpha^U[D_p(w)] \leq T_{req},\ \forall p \in \mathcal{P}$. Hence $\mathbf{z}_\alpha[D_p(w)] \leq T_{req} \iff w \in S$. Thus$S_U \subseteq S$. □

Hence, checking the satisfiability of an upper bound on the α-percentile of a path delay for a particular assignment of gate sizes, w guarantees that w is in the original feasible region. We now compute an upper bound on the $\mathbf{z}_\alpha[D_p]$. It is clear that

$$C_{ij} = \rho_{ij}\sigma(d_i)\sigma(d_j) \leq \sigma(d_i)\sigma(d_j) \tag{26}$$

Thus

$$\left(\sum_{i \in p} \sigma^2(d_i) + \sum_{\substack{i,j \\ i \neq j}} C_{ij} \right)^{\frac{1}{2}} \leq \sum_{i \in p} \sigma(d_i) \tag{27}$$

Hence, the upper bound on the α-percentile of the delay can be written as

$$\mathbf{z}_\alpha^U[D_p] = \sum_{i \in p} \mu(d_i) + z_\alpha \left(\sum_{i \in p} \sigma(d_i) \right) \tag{28}$$

In order to check the feasibility of a particular gate size assignment, we need to check whether the path having largest $\mathbf{z}_\alpha^U[D_p]$ satisfies the constraint. This can be done by assuming that every gate i in the circuit DAG has a *fixed* delay of $\mu(d_i) + z_\alpha\sigma(d_i)$ (using Equation 28) and using Dijkstra's algorithm (the complexity of this is $O(|N|+|E|)$). However, replacing the α-percentile of every path by its upper bound in the constraints results in a reduction in the feasible region of the optimization problem. Hence it might be possible that the optimal solution does not lie in the new feasible region. Thus, this transformation results in considerable reduction in the complexity at the cost of slight degradation in the quality of the solution. The quality of the solution can be improved by increasing the required time in the constraint.

4. OPTIMIZATION METHODOLOGY

Under the proposed leakage and delay models, the optimization problem is a *multivariable* non-linear optimization problem. However, when the delay constraints are replaced by their upper bounds, the problem becomes a Geometric Programming problem [16, 2]. This is of much significance as a particular class of geometric programs (where the objective and constraints are *posynomials*) can be transformed into convex optimization problems. For convex optimization problems an exact solution can be found as locally optimal solution is also the global optimum.

A posynomial function f of variable $\mathbf{x} \in \mathbb{R}^{+n}$ has the form

$$f(\mathbf{x}) = \sum_j \beta_j \prod_{i=1}^{n} x_i^{\alpha_{ij}} \tag{29}$$

where $\alpha_{ij} \in \mathbb{R}$ and $\beta_j \in \mathbb{R}^+$. The significance of posynomials comes from the fact that they can be transformed into convex functions through the transformation, $x_i = e^{z_i},\ \forall i = 1, \ldots, n$. For the proposed leakage model, the objective function of the optimization problem (3) is a posynomial only if $\lambda \geq 0.5$ (because the coefficient of one of the terms in the variance of total leakage becomes negative). Thus the optimization was performed for different values of λ such that $\lambda \geq 0.5$. We now briefly describe the convex optimization algorithm used for solving our problem.

4.1 Optimization Algorithm

This paper uses the convex optimization algorithm used in [19]. The algorithm works by successively reducing the problem region by introducing cutting planes in every iteration. The cutting planes (or hyperplanes) are obtained by conditions on the gradient of the objective functions and that of the constraints. The cutting planes are chosen such that they guarantee the presence of the optimal solution in the problem region of the next iteration. Let $\mathbf{x} \in \mathbb{R}^{+n}$ be the decision variable, $f(\mathbf{x})$ be the convex objective function and $g_i(\mathbf{x}) \leq T_{req},\ i = 1, .., n$ be the convex constraints. Let S be the feasible set defined by $\{\mathbf{x} : g_i(\mathbf{x}) \leq T_{req}\}$ and $\mathbf{x}^* \in S$ be the optimal solution. Initially, the solution space is determined by the polytope defined by the set $\{\mathbf{x} : \mathbf{x}_L \leq \mathbf{x} \leq \mathbf{x}_U\}$, where $\mathbf{x}_L$ and $\mathbf{x}_U$ are the minimum and maximum possible values of $\mathbf{x}$. Algorithm 1 outlines the convex optimization algorithm [19] for completeness.

Algorithm 1 Convex Optimization Algorithm

1: Find the center $\mathbf{x}_c$ of the current polytope P.
2: If $\mathbf{x}_c \notin S$, find the gradient $\nabla g_k(\mathbf{x})$ of the constraint having the largest value at $\mathbf{x}_c$. Goto Step 3.
3: Insert a hyperplane of the form $\mathbf{c}^T\mathbf{x} \geq \beta = \mathbf{c}^T\mathbf{x}_c$, where $\mathbf{c} = -[\nabla g_k(\mathbf{x})]^T$, update P.
4: If $\mathbf{x}_c \in S$, compute $\mathbf{c} = -[\nabla f(\mathbf{x})]^T$ and insert the hyperplane of the form $\mathbf{c}^T\mathbf{x} \geq \beta = \mathbf{c}^T\mathbf{x}_c$, update P.
5: If the size of the polytope P is less than a user specified limit ϵ, stop. Otherwise goto Step 1.

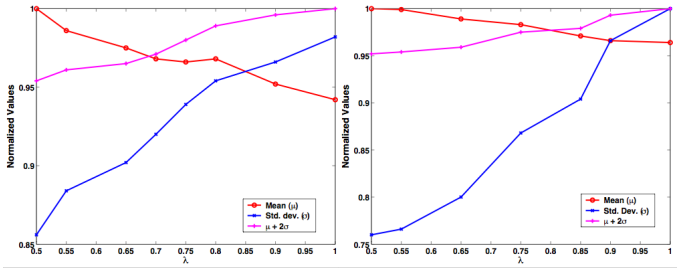
The center of the polytope in Step 1 is obtained by minimizing a log-barrier function. The complexity of computing the center has been shown to be $O(|N|^{2.5})$ [19]. The constraints in our problem are $\mathbf{z}_\alpha^U[D_p] \leq T_{req}$.

5. EXPERIMENTAL RESULTS

The proposed optimization technique was implemented in C++. The tool takes in a *blif* file as well as the parameters of the leakage and delay models of the gates. The optimization methodology was tested on ISCAS'85 benchmark circuits. The parameters of the models were fitted to correspond to 0.13μm technology. In the following sections we study the effect of various parameters of the optimization problem such as λ, T_{req} and α-percentile on the quality of the obtained solution.

5.1 Effect of varying λ

With all other parameters (T_{req}, z_α) fixed, the parameter λ varies the relative weight of mean(μ) and standard deviation (σ) in the optimization problem. Figure 2 show that for circuits C432 and C499, as λ increases from 0.5 to 1.0, the μ of the leakage decreases monotonically, whereas the σ increases monotonically. This shows that μ and σ are infact conflicting objectives in the case of leakage minimization. Hence, neither one of them can be neglected while performing the optimization. The main cause of this conflicting behavior is that the variations increase as the dimensions decrease. Since the expected leakage is a linear function of the gate sizes, decreasing the expected leakage reduces the sizes of the gates. This increases the variance of the sizes and hence the variance of the leakage power.

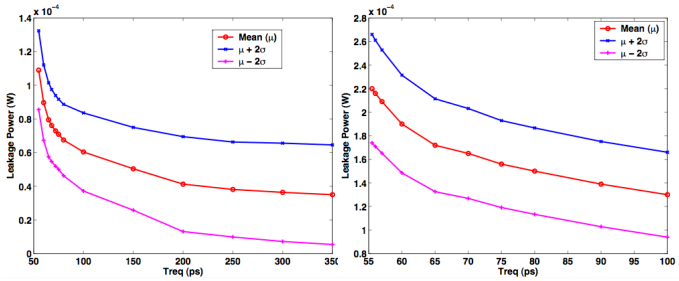


(a) (b)

Figure 2: μ-σ trade-off for: (a) C432, and (b) C499

5.2 Effect of varying T_{req}

This section discusses the power-delay trade-off for a circuit. From Figure 3, we see that as the delay constraint is tightened, the expected leakage starts to increase rapidly. This can be attributed to the fact that delay minimization is achieved by assigning larger sizes to the gates. The sizes of large gates vary by small amount. Hence the variance of the leakage as well as delay decreases. Table 2 shows that



(a) (b)

Figure 3: Power-Delay curves: (a) C432, and (b) C880

for a circuit tuned for performance, a significant power savings can be achieved by relaxing the delay constraint by only a small amount. From column 3 and 4, we see that an X amount of increase in delay can contribute to more than $2X$ reduction in the leakage.

Circuit	$T_{req}(ps)$	$\Delta T_{req}(\%)$	$\Delta \mu_{Power}(\%)$
C432	55	8.33	18.00
	60	7.69	11.36
	65	4.41	4.28
C880	55.5	0.89	2.29
	56.0	1.75	2.20
	65.0	5.00	9.83

Table 1: Power-Delay Trade-off

5.3 Effect of varying z_α

A significant amount of pessimism associated with worst case analysis can be removed by incorporating statistical design techniques. Figure 4 shows the probability distributions of an optimized circuit for different values of z_α. While considering a gate delay model such as $\mu + z_\alpha\sigma$, a higher value of z_α (3 or 4) corresponds to a more pessimistic delay model. As can be seen from Figure 4 and Table 2, use of a worst case model causes the optimization to satisfy an extremely pessimistic delay constraint which leads to a design having unusually high leakage power. This gives an impression that even more optimization is required to bring down the power. The use of statistical models and optimization techniques removes this pessimism to provide the designer a much better estimate of the expected performance. The runtime of the

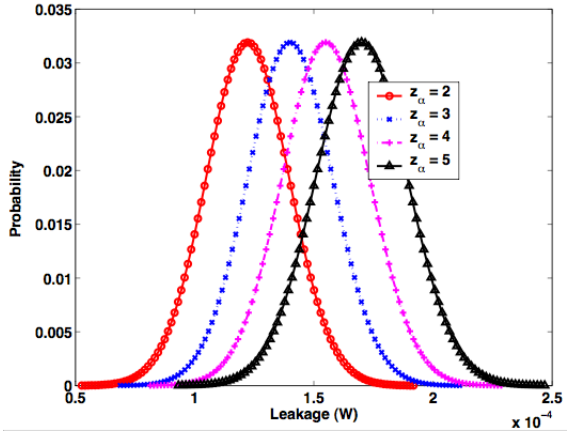


Figure 4: Effects of using a worst case model

	C432		C880	
z_α	Mean	Std. dev	Mean	Std. dev
1	49.7	12.1	97.6	18.7
2	60.4	11.6	122.2	17.4
3	71.8	10.5	139.6	17.8
4	76.9	10.4	155.4	18.4
5	87.0	10.4	170.0	19.2
6	94.3	10.9	182.9	20.2

Table 2: A pessimistic delay model causes an increase in the optimized circuit's leakage (all values in (μW))

optimization process for various benchmark circuits on an Intel Pentium IV-1.7GHz with 512MB RAM are outlined in Table 3.

6. CONCLUSIONS

We presented a novel gate sizing technique for statistical optimization of leakage power subject to constraints on

Circuit	Gates	runtime(min.)
C17	7	0.01
C432	160	3.3
C499	202	4.8
C880	383	12.2
C1355	546	38
C1908	880	65
C2670	1193	112
C3540	1660	298

Table 3: Gate sizes and runtimes of ISCAS'85 benchmark circuits

some α-percentile of the circuit delay. We derived models for the leakage power in presence of variations in the gate sizes. A particular subclass of the formulated optimization problem was identified as being convex optimization problem. This problem was then solved exactly to obtain a low leakage implementation of the circuit. This technique provides sufficient freedom (by choosing the parameters of the optimization) to the designer to fine tune it for desired performance. The technique removes the pessimism associated with the worst case design to provide a better estimate of the expected performance.

APPENDIX

The inequality in Equation 30 holds irrespective of the probability distribution of random variable w [28].

$$\left| E\left[\frac{1}{w}\right] - \frac{1}{E[w]} \right| \leq \left| \frac{1}{w} \right|_{max} \frac{\sigma_w^2}{E[w]^2} \tag{30}$$

According to the ITRS (International Technology Roadmap for Semiconductors) the maximum variation in the CD is less than 30%. Hence,

$$3\sigma_w \leq 0.30(E[w]) \tag{31}$$

$$\Rightarrow \quad \sigma_w^2 \leq 0.01(E[w]^2) \tag{32}$$

From our model, $E[w] = w_{so}$, where w_{so} is the specified size of the gate. Hence from Equation (30) and (31), we have

$$\left| E\left[\frac{1}{w}\right] - \frac{1}{E[w]} \right| \quad \leq \quad \left| \frac{1}{w} \right|_{max} \cdot 0.01 \cdot \frac{w_{so}^2}{w_{so}^2} = 0.01 \left| \frac{1}{w} \right|_{max} \tag{33}$$

Jensen's inequality states that if f is a convex function of random variable x, then

$$E[f(x)] \leq f(E[x]) \tag{34}$$

Since $1/w$ is a convex function, Equation (33) becomes

$$E\left[\frac{1}{w}\right] - \frac{1}{w_{so}} \leq \frac{0.01}{w_{min}} \tag{35}$$

where w_{min} is the minimum value of the width that can be specified by the designer. The average gate size of the gates in the circuit is around 2-4 times the minimum size (only the gates along the critical path have big sizes, other gates have small sizes). Hence $w_{so} \sim 4w_{min}$. Hence Equation (35) becomes

$$E\left[1/w\right] \leq \frac{1.04}{w_{so}} = \frac{1.04}{E[w]} \tag{36}$$

Thus, the error in this approximation is less than 4%.

1. REFERENCES

[1] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De. Parameteric variations and impact on circuits and microarchitecture. In *Proc. DAC*, 2003.

[2] S. Boyd, S. J. Kim, L. Vandenberghe, and A. Hassibi. A tutorial on geometric programming. Technical report, www.stanford.edu/∼boyd/gp_tutorial.html, 2004.

[3] C. Chen and M. Sarrafzadeh. Simultaneous voltage scaling and gate sizing for low-power design. *Trans. on CAS-II: Analog and Digital Signal Processing*, 49(6):400–408, 2002.

[4] S. W. Director et al. Optimization of parametric yield: A tutorial. In *Proc. of CICC*, pages 3.1.1–8, 1992.

[5] J. Fishburn and A. Dunlop. TILOS: a posynomial programming approach to transistor sizing. In *Proc. ICCAD*, pages 326–328, 1985.

[6] M. Ketkar and S. S. Sapatnekar. Standby power optimization via transistor sizing and dual threshold voltage assignment. In *Proc. of ICCAD*, pages 375 – 378, 2002.

[7] D. Lee, H. Deogun, D. Blaauw, and D. Sylvester. Simultaneous state, vt and tox assignment for total standby power minimization. In *Proc. of DATE*, 2004.

[8] C. Long and L. He. Distributed sleep transistors network for power reduction. In *Proc of DAC*, pages 181 – 186, 2003.

[9] V. Mehrotra, S. Nassif, D. Boning, and J. Chung. Modeling the effects of manufacturing variation on high-speed microprocessor interconnect performance. In *International Electronic Devices Meeting*, pages 767–770. IEEE, Dec 1998.

[10] V. Mehrotra et al. A methodology for modeling the effects of systematic within-die interconnect and device variation on circuit performance. In *Proc. of DAC*, pages 172–175, 2000.

[11] S. Mukhopadhyay, A. Raychowdhury, and K. Roy. Accurate estimation of total leakage current in scaled CMOS logic circuits based on compact current modeling. In *Proc. of DAC*, pages 169–174, 2003.

[12] S. Narendra et al. Full-chip subthreshold leakage power prediction and reduction techniques for sub-0.18-μ/m CMOS. *Journal of Solid-State Circuits*, 39(2):501–510, Feb 2004.

[13] S. R. Nassif. Modeling and analysis of manufacturing variations. In *IEEE Conf. on Custom Integrated Circuits*, pages 223–228, 2001.

[14] C. Neau and K. Roy. Optimal body bias selection for leakage Improvement and Process Compensation over different technology generations. In *ISLPED*, pages 116–121, 2003.

[15] M. Orshansky, L. Milor, P. Chang, K. Keutzer, and C. Hu. Impact of spatial intrachip gate length variability on the performance of high-speed digital circuits. In *IEEE Transactions on CAD*, volume 21, May 2002.

[16] E. L. Peterson. Geometric programming. *SIAM Review*, 18(1):1–51, Jan 1976.

[17] S. Raj, S. Vrudhula, and J. M. Wang. A methodology to improve timing yield in the presence of process variations. In *Proc. of DAC*, pages 448–453, 2004.

[18] R. Rao, A. Devgan, D. Blaauw, and D. Sylvester. Parametric yield estimation considering leakage variability. In *Proc. of DAC*, pages 442–447, 2004.

[19] 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. *Trans. on CAD*, 12(11):1621–1634, Nov 1993.

[20] S. Sirichotiyakul et al. Stand-by power minimization through simultaneous threshold voltage selection and circuit sizing. In *Proc. of DAC*, pages 436 – 441, 1999.

[21] A. Srivastava, D. Sylvester, and D. Blaauww. Power minimization using simultaneous gate sizing, dual-vdd and dual-vth assignment. In *Proc. of DAC*, pages 783–787, 2004.

[22] A. Srivastava et al. Statistical optimization of leakage power considering process variations using dual-vth and sizing. In *Proc. of DAC*, pages 773–778, 2004.

[23] B. E. Stine, D. S. Boning, and J. E. Chung. Analysis and Decomposition of Spatial Variation in IC processes and devices. *IEEE Trans. on Sem. Manuf.*, 10(1):24–41, Feb 1997.

[24] B. E. Stine et al. Simulating the Impact of Pattern-Dependent Poly-CD variation on circuit performance. *IEEE Trans. on Sem. Man.*, 11(4):552–556, November 1998.

[25] M. A. Styblinski. Statistical design centering approach to minimax circuit design. In *Proc. of ISCAS*, pages 697–700, 1989.

[26] T. Tugbawa et al. A mathematical model of pattern dependencies in Cu CMP processes. In *Proc. CMP Symp. Electrochem. Soc. Meeting*, pages 605–615, 1999.

[27] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, and S. Narayan. First-order incremental block-based statistical timing analysis. In *Proc. of DAC*, pages 331–336, 2004.

[28] R. von Mises. *Mathematical Theory of Probability and Statistics*. Academic Press, 1964.

[29] Q. Wang and S. Vrudhula. Algorithms for minimizing standby power in deep submicrometer, dual-vt cmos circuits. *Trans. on CAD*, 21(3):306–318, 2002.