Joint Design-Time and Post-Silicon Minimization of Parametric Yield Loss using Adjustable Robust Optimization

Murari Mani, Ashish K. Singh, and Michael Orshansky

Department of Electrical and Computer Engineering, University of Texas at Austin

mani@ece.utexas.edu

ABSTRACT

Parametric yield loss due to variability can be effectively reduced by both design-time optimization strategies and by adjusting circuit parameters to the realizations of variable parameters. The two levels of tuning operate within a single variability budget, and because their effectiveness depends on the magnitude and the spatial structure of variability their joint co-optimization is required. In this paper we develop a formal optimization algorithm for such co-optimization and link it to the control and measurement overhead via the formal notions of measurement and control complexity.

We describe an optimization strategy that unifies design-time gate-level sizing and post-silicon adaptation using adaptive body bias at the chip level. The statistical formulation utilizes adjustable robust linear programming to derive the optimal policy for assigning body bias once the uncertain variables, such as gate length and threshold voltage, are known. Computational tractability is achieved by restricting optimal body bias selection policy to be an affine function of uncertain variables. We demonstrate good run-time and show that 5-35% savings in leakage power across the benchmark circuits are possible. Dependence of results on measurement and control complexity is studied and points of diminishing returns for both metrics are identified.

Categories and Subject Descriptors

B.7.2 [**Integrated Circuits**]: Logic Design– *optimization*

General Terms

Algorithms, Design, Reliability

1. INTRODUCTION

Increased variability of device parameters necessitates the development of a new generation of circuit synthesis CAD tools. In addition, the increases in variability and power consumption are closely related because of the exponential dependence of leakage on some process and environment parameters. Two fundamental paradigms are available for dealing with variability: statistical design (optimization at design time) and post-silicon adaptivity (on-line tuning). To guarantee reliable circuit operation with minimal power consumption, next-generation circuit synthesis techniques for robustness must explicitly account for the availability of post-silicon adaptivity in synthesizing the circuit.

ICCAD'06, November 5-9, 2006, San Jose, CA

Two powerful and complementary strategies for reducing leakage considering variability are pre-silicon statistical design optimization and post-silicon adaptivity. There is a growing body of work on statistical circuit analysis methods [1-3] (i.e., SSTA) and statistical post-synthesis optimization [4-6], including sizing and dual-threshold voltage assignment algorithms. These tools show promise in reducing parametric yield loss, or alternatively, reducing power consumption while maintaining high yield: in some cases, a 25% reduction in power is gained at the cost of 5% timing yield loss. The growing magnitude and complexity of uncertainty is bound to make post-synthesis optimization techniques insufficient in guaranteeing reliable circuit operation with reasonable parametric yield.

Post-silicon design adaptivity, or tuning, currently includes several techniques; the primary ones being adaptive body biasing (ABB) and adaptive supply voltage (ASV). ABB uses the body effect to modulate the threshold voltages of transistors, thereby controlling leakage and performance [7-10]. ASV raises the power supply (V_{dd}) for slow (low-leakage) dies, and lowers it for fast (high-leakage) dies, ensuring better overall yield [11]. It relies on the roughly cubic dependence of leakage power on V_{dd} in CMOS circuits (also impacting dynamic power quadratically). In the future, a larger palette of tuning tools is likely to emerge: recently, an adaptive-size tapered Pareto buffer was designed with control facilitated via a tri-state buffer [12].

A widespread industrial adoption of adaptive techniques is not yet possible for two reasons. One is that designers do not have the tools to help them decide whether, and how much, adaptive circuitry is needed, or what type of post-silicon tuning technique will be most appropriate. The availability of both design-time (pre-silicon) optimization and post-silicon adaptivity leads to a rich optimization space in which coordination between the two levels is required. Sizing can be used to upsize the gates beyond the need of a nominal design to achieve higher timing yield, but with increased power. Alternatively, the adaptivity of threshold voltage can be used to tighten the speed distribution to improve yield. Depending on the magnitude and the spatial structure of variability, the two approaches will have different cost-effectiveness, i.e., they will be characterized by different Pareto curves in the space of design objectives.

Algorithmically, future robust circuit synthesis can be conceptualized as a two-stage optimization problem, with additional second-stage tuning available upon the realization of uncertain variables. In this paper an efficient formulation is proposed using the theory of adjustable optimization. This optimization paradigm presumes that the decision-maker has a chance to update his optimization strategy upon learning additional information. If the objective function is linear in the decision variables, then, under the conditions that the uncertainty

sets are affine functions of some parameters, the optimal policy for the second-stage decisions can be computed efficiently. Two stage stochastic optimization problems are difficult to solve [13]. However, recent developments in the theory of convex programming have enabled the solution of robust versions of linear and quadratic programs, which can be expressed as second-order conic programs or semi-definite programs [14, 15]. Extending these solution methods to adjustable robust programming has been demonstrated in [16]. In this paper, we build upon this work to develop an efficient solution to the post silicon optimization problem under variability.

The problem is formulated in the following way. The first-stage (design-time) power-delay optimization is done via sizing, and second-stage (post-silicon) optimization is achieved by body bias tuning. The second stage decision variables are represented as affine function of parameter uncertainty. The solution to this optimization problem is a design time decision (size of gates in the circuit) and an optimal policy that prescribes the amount of bias depending on the realizations of uncertain variables (e.g. gate length, V_{th} on a specific chip). Initial experiments prove that the optimal synergy between design-time and post-silicon optimization depends on the amount and structure of variability. If variability is highly spatially correlated within the chip, design-time optimization will be ineffective and may even lead to large yield losses. On the other hand, with the increase of intra-chip variability, the effectiveness of post-silicon adaptivity decreases. Three measures of complexity that parameterize the solution and the optimality of this problem are introduced by us: the control complexity (the granularity of control), the measurement complexity (the granularity of the monitoring and sensing circuitry), and the parameter complexity (a measure of how spatially uncorrelated the process variable is). Using these metrics, formal quantitative trade-offs between design-time and post-silicon adaptivity can be identified. Such capability will also be useful for the analysis and development of the fine-granular control structures, e.g. for determining the spatial granularity.

The rest of the paper is organized as follows. Section 2 motivates the need for joint co-optimization between design-time and post-silicon optimization. Section 3 presents the leakage and delay models used. The details of the algorithm are presented in Section 4 followed by the results and analyses in Section 5.

2. DESIGN-TIME / POST-SILICON CO-OPTIMIZATION: MOTIVATION AND CHALLENGES

The central problem of statistical optimization methods is reducing the dual parametric yield loss due to power and timing constraints. This is because power consumption has become the yield-limiting factor, indirectly affecting the achievable maximum clock frequency [17]. In the absence of substantial leakage power, parametric yield is determined by the maximum possible clock frequency. When realistic leakage power numbers for current CMOS technologies are added, the total power starts approaching the power limit determined by the cooling and packaging considerations. Crucially, the exponential dependence of leakage on process spread will mean that the total power will cross the cooling (power) limit well below the maximum possible chip frequency, since chips operating at higher frequencies have exponentially higher leakage power consumption. Due to the inverse correlation between speed and leakage, yield is limited both by slower chips and chips that are too fast, because they are too leaky.

The fundamental limitation of design-time methods is that they impose an overhead on *each* instance of the fabricated chip since they intrinsically lack the ability to "react" to the actual conditions on the chip. For example, when using sizing for timing optimization they impose a fixed area overhead that may be wasteful on some instances of the ICs that would meet timing even with smaller driver sizes. Having an adjustable-width driver would be ideal, since it could ensure meeting constraints with the minimum overhead for each chip.

The problem that we address in this paper is how to perform design-time circuit optimization and post-silicon tuning jointly. Why should these two steps be coordinated, i.e., why do we need joint co-optimization? The two methods operate from different viewpoints: in design-time optimization a decision (e.g., sizing) must be made before the realization of uncertainty (gate length), while in post-silicon tuning of the decision (the value of bias to apply) is made after the realization of uncertainty, i.e., when the chip's physical properties have been determined during manufacturing.

However, the two paradigms operate within a single budget of uncertainty, and thus meeting constraints can be achieved by both methods. But their cost-effectiveness depends on specific conditions, such as the spatial correlation of process variability, the granularity of adaptivity that can be implemented, and the magnitude of leakage power in comparison with the switching power. The objective of this paper is to develop formal means and optimization methods that will allow joint optimization. The specific optimization strategy will jointly consider the amount of variability and cost-effectiveness of power reduction strategies, to derive *a policy* that will guide post-silicon tuning, as well as make the first-phase design decisions. This will allow to optimally partition the design space between these levels of hierarchy.

Formally, the objective of the algorithm we develop is to minimize the expected value of leakage power under a given delay constraint T at a given yield α :

$$\min \; E_{leak} \; s.t. P(D \le T) \ge \alpha$$

This formulation is generic and, different specific optimization mechanisms can be studied. In this paper we focus on sizing and adaptive body bias for threshold control at the chip level, with only a small number of partitions of the chip into individually tunable clusters. The widely different spatial scales involved in this problem are of some interest and are actively explored. In the above formulation, the objective function and the constraints depend on both the design time optimization variables (sizes) and the post silicon decision variables (body biases). The problem can be formally viewed as a two-phase optimization under uncertainty with recourse. The key contribution of our approach is the derivation of the optimal policy for body biasing as an affine function of the realizations of the uncertain parameters (gate length, L , and threshold voltage, V_{th}). The solution to the above optimization problem therefore yields the sizes for the gates in the circuit and an optimal body bias policy.

3. GATE AND CIRCUIT MODELING

Adjusting the circuit properties to manufacturing conditions can be achieved by several techniques, including adaptive buffer

sizing, adaptive body biasing, and adaptive supply voltage biasing. Because joint timing-leakage optimization is of primary concern, adaptive body bias may be the most useful tool. It has been demonstrated [7, 11] that body biasing can be employed as an extremely effective knob to perform post silicon optimization and performance tuning by reducing the leakage for those dies that violate power constraints and increasing the frequency of those dies that do not meet delay specs.

The adaptive body bias technique exploits the dependency of the threshold voltage of a MOSFET device on its source-to-body voltage to achieve dynamic tuning of its delay and leakage power. For an NMOS device, the threshold voltage can be expressed as [18]:

$$V_{th} = V_{th0} + \gamma(\sqrt{V_{SB} + 2\phi_f} - \sqrt{2\phi_f})$$

where V_{th0} is the threshold voltage of the device with zero body bias, γ is the body bias coefficient, and ϕ_f is the Fermi potential. Decreasing the source potential relative to the body of an N-channel device, translates to a negative V_{SB}, and decreases the threshold voltage. This technique, known as forward body biasing (FBB) reduces the delay of the gate at the expense of leakage power. On the other hand, application of reverse body bias (RBB) by applying a positive V_{SB} causes the threshold voltage of the device to increase. RBB is thus very effective in reducing the leakage power consumption [1].

The need to setup a rigorous statistical optimization problem under uncertainty requires us to use approximate, linearized delay models, such as a piecewise delay of [19]. Let the gate delay be represented as $d_i = \overline{d}_i + \Delta d_i$, where $\overline{d}_i$ is the nominal gate delay and Δd_i is the term representing the variability in delay. The dependence of nominal gate delay on gate sizes can be described by the piecewise linear equations:

$$\overline{d}_i = a_{i1}^l - a_{i2}^l w_i + a_{i3}^l \sum w_k \; \forall\, l \in [1, m] \tag{1}$$

where m is the number of fitting regions l and a_is are the fitting coefficients. This model captures the dependence of delay on the size of the gate width w_i, and its load $\sum w_k$. The accuracy of the approximation is reasonable: the average error is less than 5% for $m = 3$. The size range considered is 1-8x of the minimum size gate.

Analytical models are used to relate the impact of variability sources on power and delay. The variability is assumed to come from two major sources. Transistor gate length (L) exhibits strong lithography induced variability. Threshold voltage (V_{th}) variation due to oxide thickness and dose variation is also taken into account. The impact of L on V_{th} due to drain-induced barrier lowering is predicted by the device model directly, which permits modeling L and V_{th} as independent random variables. Both L and V_{th} are assumed to follow the normal distribution. An additive statistical model that decomposes the variability, of both L and V_{th}, into the global (chip-to-chip) and local (intra-chip) uncorrelated variability components is used. For gate length:

$$L = L_0 + \Delta L_g + \Delta L_l \tag{2}$$

The impact of process parameter variability on gate delay is captured using a first-order parametric delay model:

$$\Delta d \cong S_1 \Delta L + S_2 \Delta V_{th} + S_3 \Delta V_{SB} \tag{3}$$

where ΔL and ΔV_{th} are the parameter deviations and ΔV_{SB} is the applied body bias. The sensitivities are the first-order derivatives of delay with respect to the specific variable (L, V_{th}, V_{SB}).

Using a modeling approach similar to [20], the subthreshold leakage current of a gate is expressed as an exponential function of the random parameters as:

$$I = I_o \cdot \exp(a\Delta L + b\Delta V_{th} + c\Delta V_{SB}) \tag{4}$$

where I_o is the nominal value of leakage per unit width. We obtain a good fit using this model (Figure 1), the *rms* error being ~8%. For a circuit block the expression for leakage can be expressed as:

$$I_{tot} = \sum_i \beta_i \cdot w_i \cdot \exp(a_i \Delta L_i + b_i \Delta V_{th,i} + c_i \Delta V_{SB})$$

where the nominal gate leakage is $I_{0,i} = \beta_i \cdot w_i$.

Following [21], we assume that the impact of random component of variation on chip-level leakage value can be captured by a constant multiplier that we take to modify the value of β_i, in the above expression.

The essence of adjustable optimization framework is that the variable that is allowed to be tuned is not determined arbitrarily but is dependent in some way on the realizations of uncertain variables. As was mentioned before and will be justified in the next section, a computationally tractable solution to a statistical adjustable problem requires that ΔV_{SB} be an affine function of uncertain parameters, L and V_{th}:

$$\Delta V_{SB} = \pi_0 + \pi_1 \Delta L_g + \pi_2 \Delta V_{th,g} \tag{5}$$

Here, the coefficients π_0, π_1 and π_2 are to be determined in the process of optimization. Such a parameterization is physically equivalent to compensating for the variation in leakage due to L and V_{th}, by applying body bias [22]. Though, the value of body bias is not a random variable, based on (5), it can be treated mathematically as one. With that observation, let us define:

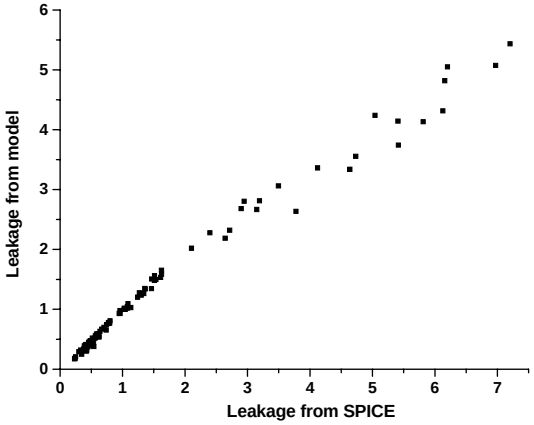

Figure 1. Comparison of the normalized leakage of inverter predicted by SPICE and the analytical leakage model.

$X_i = N(\mu_i, \sigma_i^2) = a_i \Delta L_g + b_i \Delta V_{th,g} + c_i \Delta V_{SB}$

The mean and variance of a lognormal $Y = e^X$ in terms of the mean and variance of the normal random variable $X = N(\mu, \sigma^2)$ are [23]:

$$E(Y) = \exp(\mu + \frac{\sigma^2}{2}) \quad (6)$$

$$Var(Y) = \exp[2(\mu + \sigma^2)] - \exp(2\mu + \sigma^2) \quad (7)$$

Observing that :

$E(I_{tot}) = \sum_i E(\beta_i \cdot w_i \cdot \exp(a_i \Delta L_g + b_i \Delta V_{th,g} + c_i \Delta V_{SB}))$

and $\mu_i = E(a_i \Delta L_g + b_i \Delta V_{th,g} + c_i \Delta V_{SB}) = \pi_o$

we can write the expected value of total block leakage as:

$$E(I_{tot}) = \sum_i \beta_i \cdot w_i \cdot \exp(\pi_o + \sigma_i^2 / 2) \quad (8)$$

4. DESIGN TIME / POST SILICON CO-OPTIMIZATION ADAPTABLE ROBUST OPTIMIZATION

In the optimization strategy we develop, the optimal body bias is determined after the realization of uncertainty of the process parameters. On-chip measurements are used to measure the actual parameter values and their deviations from nominal values. Then, the policy derived during optimization can be used to choose an optimal forward or reverse body bias. RBB can be applied to reduce yield loss in the high frequency (high leakage bins), and can be used with FBB to tighten the distribution at the low frequency bins.

4.1 Adaptable Robust Optimization

First we introduce the theoretical foundation for robust adjustable optimization. We use robust optimization as the bedrock of our strategy. A robust LP can be defined as the problem of minimizing the worst-case realization of a linear objective over a set of linear constraints with uncertain coefficients [24]:

$$\min\{ \sup_{\zeta \equiv [A,b,c] \in Z} (c^T x) : Ax \le b \ \ \forall \zeta \equiv [A,b,c] \in Z\} \quad (9)$$

Here the uncertainty in the matrix coefficients is represented as $\zeta \equiv [A,b,c]$ varying in the nonempty compact convex uncertainty set Z .

The above problem requires all decisions to be made prior to the actual realization of the uncertain parameters. However, in many real-life cases not all the decisions can be made simultaneously: only some variables may become known earlier. In this case, the remaining decision variables can be adjusted to the realizations of uncertain data. It is obvious that if the opportunity to adjusting some variables is given, the optimal solution will be better (or at least, no worse) than for the problem above. Problems with similar structure have been known as multi-stage stochastic problems with recourse. However, robust problems are not stochastic problems, and when certain conditions are imposed on the uncertainty set, demonstrate superior computational properties.

We can re-write the problem of (9) in terms of the non-adjustable variables u and the adjustable variables v . This leads to the adjustable robust problem:

$$\min\{c^T \begin{pmatrix} u \\ v \end{pmatrix} : \forall(\zeta \equiv [U,V,b,c] \in Z) \ \exists v : Uu + Vv \le b\} \quad (10)$$

In this formulation, the adjustable variables v are allowed to depend on the realization of ζ .

Still, it is shown in [16] that the general robust problem with adjustable parameters is NP-complete, unless restrictions are applied on how exactly the adjustable variables tune themselves to uncertain data. It is shown that a computationally feasible adjustable robust linear problem can be achieved if the adjustable variables are constrained to be affine functions of the uncertain variables. This is equivalent to:

$v = w + W\zeta$

From this we see that the adjustable variables can be tuned once the realization of uncertain data is known. However, if we are to be able to identify an optimal policy and do that computationally efficiently, the dependency cannot have general form, but must be constrained. This ultimately leads to the affinely adjustable robust linear program

$$\min\{c^T \begin{pmatrix} u \\ v \end{pmatrix} : Uu + V(w + W\zeta)v \le b \ \forall \zeta \equiv [U,V,b,c] \in Z\} \quad (11)$$

In particular, for uncertainty sets specified using linear or second-order cone constraints, the above problem can be reformulated as an LP or a second-order conic program respectively [16].

4.2 Co-Optimization: Problem Formulation

We now map our design-time and post-silicon tuning problem into a robust adjustable linear program. Our objective in formulating the problem is to set up a robust linear program with adjustable parameters. Robust programs have been recently used for several CAD problems [5], and are very efficient.

The task of co-optimization is effectively finding the solution to a two-stage optimization problem with recourse. Denoting column vectors by boldface letters, we formulate the problem as that of minimizing the overall expected leakage power (or current) with expectation being taken over the population of manufactured chips while satisfying timing constraints under a statistical timing model:

$$\min E(I_{tot}) \quad s.t. P(D(\boldsymbol{w}, \Delta V_{SB}) \le T) \ge \alpha \quad (12)$$

In this formulation the objective and constraint functions are dependent both on design-time variables (gate sizes) and post-silicon optimization variables (ΔV_{SB}).

We begin by writing the expression for mean leakage as:

$E(I_{tot}) = \boldsymbol{g}^T \boldsymbol{w}$

where $\boldsymbol{g}$ is an $N \times 1$ vector with entries $g_i = \beta_i \exp(\pi_o + \sigma_i^2 / 2)$. The objective function is thus linear in the gate sizes and non-linear (exponential) in ΔV_{SB} . We will deal with this by adopting a linearization approach in which we locally linearize the objective's dependence on ΔV_{SB} at the fixed value of vector of gate sizes $\boldsymbol{w}$.

Let us, for convenience form a single vector of decision variables $\boldsymbol{x} = [\boldsymbol{w} \ \Delta V_{SB}]^T$. The gate delay model introduced in the previous section can allow us to express path timing constraints in the form of:

$D(\boldsymbol{w}, \Delta V_{SB}) = a^T x$.

Now consider the probabilistic chance constraint $P(D(\boldsymbol{w}, \Delta V_{SB}) \leq T) \geq \alpha$ specified for the entire circuit. We can heuristically re-write the circuit-level probabilistic timing constraints in terms of path-based constraints. We assume that a corresponding confidence level η_j can be selected using the strategy outlined in [25]. Then, we require:

$$P(D_i(\boldsymbol{w}, \Delta V_{SB}) < T) \geq \eta_j \quad \forall j \in \Pi$$

where Π is the relevant path-set. Relying on the linear vector representation introduced above we can write:

$$P(D_i(\boldsymbol{w}, \Delta V_{SB}) \leq T) = P(\boldsymbol{a^T x} \leq T) \geq \eta_j \, \forall j \in \Pi$$

If $\boldsymbol{a}$ is distributed normally, $N(\bar{a}, \Sigma)$, the coefficients of $\boldsymbol{x}$ belong to an ellipsoidal uncertainty set [24]. Then, it can be shown that the above constraint is equivalent [14] to:

$$\bar{\boldsymbol{a}}^T \boldsymbol{x} + k_j (\boldsymbol{x}^T \Sigma \boldsymbol{x})^{1/2} \leq T \quad \forall j \in \Pi \tag{13}$$

where $k_j = \phi^{-1}(\eta_j)$ and ϕ is the cumulative distribution function (*cdf*) of the standard normal distribution. The path delay constraints of Eq. 13 represent a set of second-order conic path timing constraints [14]. Second-order conic programs are convex, and there exist extremely efficient techniques to solve SOCPs that exploit their special structure [26]. While the worst-case complexity of interior-point methods for SOCP is polynomial, for most practical instances the run-time behavior is much better. In our case, the empirically observed complexity is close to $O(N)$.

It has been shown that adjustable robust linear programs can be made computationally tractable only if the adjustable (second-stage) decision variables are *affine* functions of uncertain variables [16]. Without loss of generality, consider only the global sources of variation ΔL_g and $\Delta V_{th,g}$, and a single value of body bias ΔV_{SB} for all the gates on the chip. Then, the affine policy is given by:

$$\Delta V_{SB} = \pi_0 + \pi_1 \Delta L_g + \pi_2 \Delta V_{th,g} \tag{14}$$

This dependence can be used to express the expected value of leakage current as:

$$g_i = \beta_i \exp\left(f_{0,i}^2(\pi_0) + f_{1,i}^2(\pi_1)\sigma_{L_g}^2 + f_{2,i}^2(\pi_2)\sigma_{V_{th,g}}^2 \right) \tag{15}$$

where $f_{0,i}$, $f_{1,i}$, and $f_{2,i}$ are linear functions of π_0, π_1, and π_2 respectively.

The final robust adjustable optimization problem $ABB(\boldsymbol{w}, \boldsymbol{\pi})$ can now be expressed as:

$$\begin{aligned} &\min \ \boldsymbol{g}^T \boldsymbol{w} \\ &\bar{\boldsymbol{a}}_j^T \boldsymbol{x} + \phi^{-1}(\alpha_j)(\boldsymbol{x}^T \Sigma_j \boldsymbol{x})^{1/2} \leq T \quad \forall j \in \Pi \end{aligned} \tag{16}$$

where

$$g_i = \beta_i \exp\left(f_{0,i}^2(\pi_0) + f_{1,i}^2(\pi_1)\sigma_{L_g}^2 + f_{2,i}^2(\pi_2)\sigma_{V_{th,g}}^2 \right) \ \forall i \in [1, N]$$

Note that the original problem has now been cast as an optimization problem in π_0, π_1, and π_2 and gate widths, w_i. The solution to this problem is an optimal policy $P = (\pi_0, \pi_1, \pi_2)$ and the vector of gate width $\boldsymbol{w}$ such that the timing constraints are satisfied.

4.3 Problem Solution

To enable a computationally efficient solution, we solve the problem in (16) as a two phase optimization program. The first phase consists of solving a weighted sizing problem assuming fixed body bias and the second phase consists of solving for the body bias value assuming fixed gate size. This is performed in an iterative manner using successive approximations until the solution converges. We transform the path based formulation into a node based formulation [25] to solve the problem efficiently.

This problem is solved iteratively by computing optimal ws in the first stage and optimal πs in the second stage until the solution converges. At an iteration l the w-phase consists of solving $ABB(\boldsymbol{w}, \boldsymbol{\pi}^{(l-1)})$ to obtain $\boldsymbol{w}^{(l)}$ and the π-phase solves the problem $ABB(\boldsymbol{w}^{(l)}, \boldsymbol{\pi})$ to obtain $\boldsymbol{\pi}^{(l)}$. Initially, for $l = 0$, $\pi_j = 0 \ \forall j \in [1, k]$ corresponding to zero body bias.

Solving w-phase does not pose a problem as the objective function is linear in gate widths, $\boldsymbol{w}$ and the delay constraints are second order cones. It can therefore be solved readily as an SOCP. However, the π-phase objective is non-linear in the decision variables. To address this issue we propose to expand the objective function using a first order Taylor series. The π - phase optimization problem solved at iteration l is approximated as:

$$\begin{aligned} &\min F \\ &s.t. F \geq \pi_0 \nabla_{\pi_0}(\boldsymbol{g}^T \boldsymbol{w}) + \pi_1 \nabla_{\pi_1}(\boldsymbol{g}^T \boldsymbol{w}) + \pi_2 \nabla_{\pi_2}(\boldsymbol{g}^T \boldsymbol{w}) \\ &\bar{\boldsymbol{a}}_j^T \boldsymbol{x} + \phi^{-1}(\alpha_j)(\boldsymbol{x}^T \Sigma_j \boldsymbol{x})^{1/2} \leq T \quad \forall j \in Paths \end{aligned} \tag{17}$$

where ∇_{π_0}, ∇_{π_1} and ∇_{π_2} are the gradients computed w.r.t π_0, π_1, and π_2 respectively. The complete algorithm **optim_abb** is presented in Figure 2.

4.4 Handling of intra-chip variation

The policy described above cannot account for random parameter variation. Since the structure of the policy needs to be specified at optimization time, we need to know the number of measurements we can make on chip to account for the intra-chip random variation. Assume that we can make k_l measurements and k_v measurements of V_{th}. Assuming that we are allowed a single choice of ΔV_{SB}:

$$\Delta V_{SB} = \pi_0 + \sum_{i=1}^{k_l} \pi_i \Delta L_i + \sum_{i=1}^{k_v} \pi_{k_l+i} \Delta V_{th,i} \tag{18}$$

The notion of measurement complexity $k = k_l + k_v$ is used here to represent the amount of information we are able to obtain about the structure of variability. As we demonstrate in the results section, a higher value of k implies a lower leakage value. However, it is achieved at the cost of increased run-time and diagnostic overhead.

Similarly, we can introduce the notion of control complexity n which refers to the number of body bias values that are allowed. Control complexity reflects the degree of controllability over the body bias assignment and also the circuit overhead. It is currently assumed that the granularity of body bias assignment is at the

1. **set** $\pi_i = 0 \ \forall i \in [1,k]$
2. get Timing Target T
3. **set** $D < T$ such that $ABB(\boldsymbol{w}, \boldsymbol{\pi}^{(l-1)})$ is feasible.
4. chose delay increment δD
5. **set** $l = 1$
6. **if** $D < T$

   solve w - phase $ABB(\boldsymbol{w}, \boldsymbol{\pi}^{(l-1)})$ setting delay constraint to D.

   else

   print $\boldsymbol{w}^{(l-1)}$ and $\boldsymbol{\pi}^{(l-1)}$ as the optimal solution and stop
7. **set** $D = D + \delta D$
8. **if** $D < T$

   solve π -phase $ABB(\boldsymbol{w}^{(l)}, \boldsymbol{\pi})$ setting delay constraint to D.

   else

   print $\boldsymbol{w}^{(l)}$ and $\boldsymbol{\pi}^{(l-1)}$ as the optimal solution and stop
9. **set** $D = D + \delta D$
10. **set** $l = l + 1$ and goto step 6

Figure 2. The two phase algorithm optim_abb for post silicon optimization using ABB.

block level. This is because tuning individual gates is clearly too expensive from the physical design perspective (extra routing overhead, voltage conversion). Spatial clustering may also be used as the gates that are spatially proximate are more likely the benefit from an equal body bias assignment.

5. EXPERIMENTS AND RESULTS

We are now in a position to put together the complete design-time and post silicon co-optimization flow. We start by choosing the level of measurement complexity and control complexity. These along with the distributional information about the uncertain data are the inputs to the above algorithm. The algorithm **optim_abb** produces a set of gate sizes and an optimal policy for selecting ΔV_{SB} for the given structure of variation and the control and measurement complexity. When the chip is fabricated, the actual realizations of the uncertainty are known hence the value of body bias is determined from the policy from (18). The optimization problem was solved using the conic optimization package MOSEK [27]. The experiments were run on a 32-bit, 3.7 GHz. Intel Xeon processor with 4GB of memory. The benchmark circuits were synthesized to a cell library that was characterized for a 70 nm process using Berkeley Predictive Technology Model [28]. For NMOS (PMOS) transistors, the threshold voltage is 0.10V (-0.10V). The assumed magnitude of V_{th} and L variability is $\sigma/\mu = 8\%$ and 5% respectively. The optimal solution (sizes and policy) produced by the algorithm were evaluated using Monte Carlo analysis to estimate the expected value of leakage power by sampling from the distribution of the uncertain parameters V_{th} and L.

Three measures of complexity are used to characterize the optimality of the solution: the control complexity n which represents the granularity of control, the measurement complexity k which refers to the granularity of the monitoring and sensing circuitry, and the parameter complexity ρ, defined as the ratio $\sigma_l^2 / \sigma_{tot}^2$. Thus, ρ is a measure of how spatially uncorrelated the process variable is.

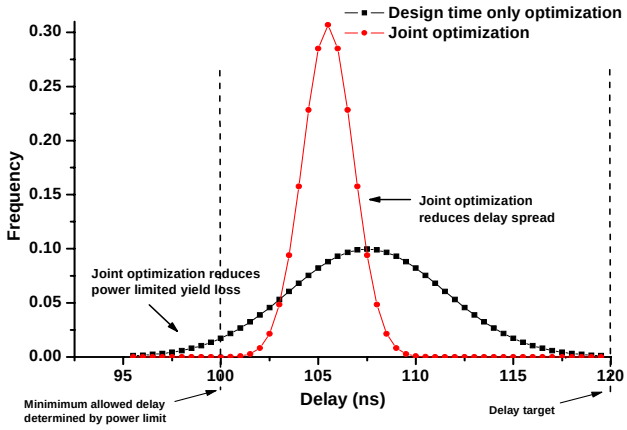

Figure 3. PDFs of delay distributions produced by design-time only optimization and joint optimization. Joint optimization is successful in tightening the distribution.

Figure 3 illustrates the effectiveness of our algorithm in reducing the spread of the circuit delay and ameliorating the problem of the dual ended squeeze on parametric yield. This is achieved by increasing the delay of faster chips by applying RBB. Since these chips have high leakage power consumption, our algorithm reduces power limited yield loss. From the Figure it can be seen that the yield is improved by about 5%. Application of FBB to slow chips serves to tighten the delay distribution further. Since the circuit is guaranteed to meet the timing yield target even for zero FBB, applying forward body bias does not improve timing yield but increases the number of chips in the higher frequency bins.

Figure 4 compares the leakage power of the circuits obtained by employing only design time optimization and the joint design time and post silicon algorithm outlined in the paper. As expected, using post silicon optimization enables a more optimal solution compared to design time only optimization. However as the complexity of variability increases, the benefit of using post silicon optimization decreases. This can be attributed to the fact that as the amount of uncorrelated variability increases, design time optimization performs better, but to utilize the adaptability provided by post silicon optimization, more measurements need to be made and more complex control system used (larger number of individually tuned clusters of logic on a chip). Therefore, increasing measurement complexity k improves the quality of the solution (reduces expected value of leakage). This is also depicted in Figure 4. However, this comes at the cost of increased run-time and diagnostic overhead. This is shown in Figure 5, which indicates that the run-time of the algorithm increases as k is increased.

Table 1 documents the results obtained across the benchmarks. All solutions were evaluated using Monte Carlo analysis. 1000 samples were generated for each random parameter. The circuits were optimized for the same delay target, which is evaluated using Monte Carlo. We observe that for a reasonable choice of measurement complexity, using our algorithm, an average saving

Table 1: **Leakage power savings obtained by the joint post-silicon and design-time optimization.**

Circuit	No. of gates	Design time optimization $E(I_{leak})$ (μW)		Joint design-time and post-silicon optimization ($k = 8$)				
		$\rho = 0.5$	$\rho = 0.8$	$E(I_{leak})$ (μW)		Leakage power savings (%)		Runtime (s)
				$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	
C432	261	328	301	246	291	25.00	3.32	8
C499	641	908	845	568	622	37.44	26.39	15.2
C880	615	560	470	388	405	30.71	13.83	12.5
C1355	685	684	603	557	595	18.57	1.33	21.1
C1908	1238	1203	1167	926	1040	23.03	10.88	31
C2670	2041	1706	1669	1405	1530	17.64	8.33	55
C3540	2582	2718	2584	2142	2473	21.19	4.30	63
C5315	3753	3801	3700	3544	3598	6.76	2.76	108
C6288	2704	2918	2902	2454	2685	15.90	7.48	132
Average savings						**21.8**	**8.73**	

of 20% savings in leakage power consumption can be obtained compared to design time only optimization. Table 1 also cites the runtimes of the algorithm. It can be seen that the run time behavior is extremely good (about 2 minutes) even for the largest benchmark circuit.

Finally, we explore the dependence of the quality of the solution obtained from post silicon optimization on the measurement complexity and control complexity. Increasing k improves the leakage power but there is a point of diminishing returns beyond which the improvement is insignificant. This is depicted in Figure 6.

Increasing the number of circuit clusters with individually adjustable threshold voltages (i.e., increasing the control complexity) improves the results of optimization, Figure 7. As with measurement complexity, this improvement in leakage power is achieved at a cost. A larger value for control complexity implies greater overhead, such as in biasing circuitry and routing.

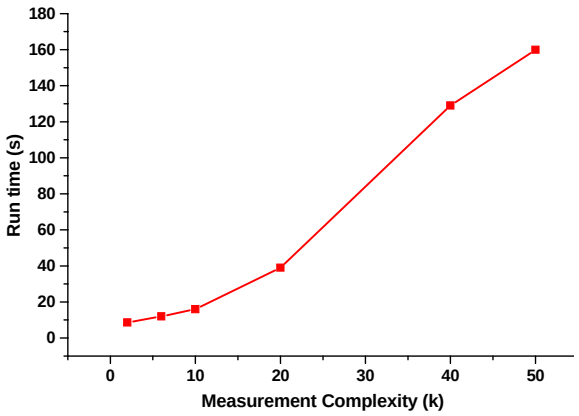

Figure 5. The runtime increases as the measurement complexity is increased, as optimal policy depends on more measurements.

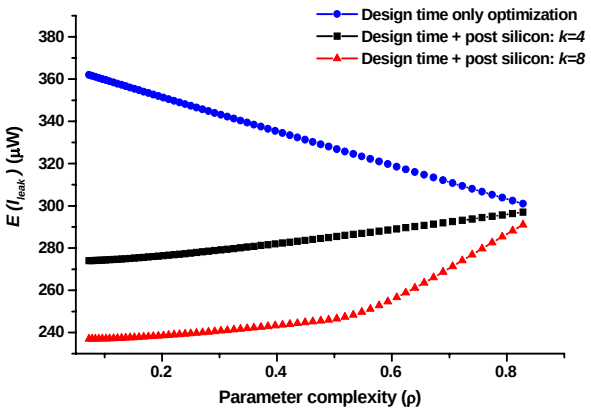

Figure 4. Comparison of design time only optimization and joint design time and post silicon optimization. Joint optimization always does better than design time only optimization.

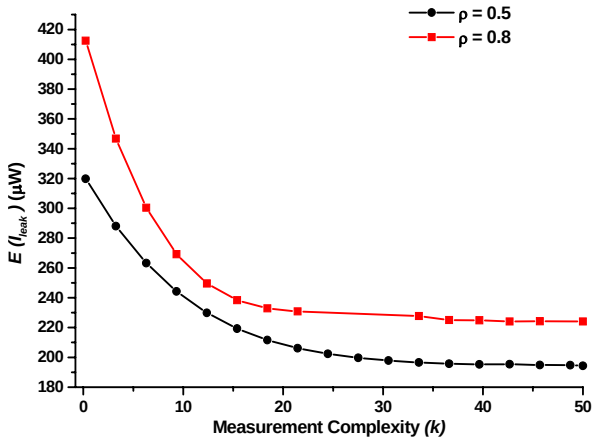

Figure 6. The expected value of leakage power decreases as we increase measurement complexity but the benefits level off for high values of *k* .

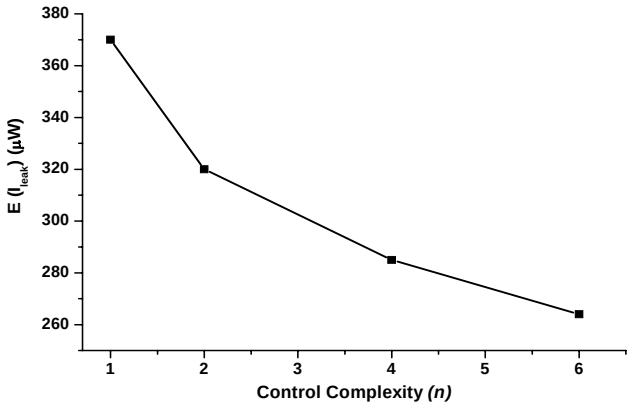

Figure 7. The expected value of leakage power decreases as we increase control complexity *n*. A larger *n* corresponds to more allowable values of body bias.

6. CONCLUSION

In this paper we have developed a theoretical foundation for joint design-time and post-silicon optimization. The problem is cast as an adjustable robust linear program and solved in a computationally efficient way. Results indicate that the designer can greatly benefit from synergistic application of design time and post silicon optimization techniques due to the ability of post silicon optimization solution to tune itself to the realization of uncertain data.

7. ACKNOWLEDGMENTS

This research was supported in part by SRC, GSRC, NSF, SUN, and Intel.

8. REFERENCES

[1] Visweswariah C. *et. al.*, "First-order incremental block-based statistical timing analysis," in *Proc. DAC*, 2004, pp. 331-336.

[2] Hongliang C. and Sapatnekar S., "Statistical timing analysis considering spatial correlations using a single PERT-like traversal," in *Proc. of ICCAD*, 2003, pp. 621 – 625.

[3] Devgan A. and Kashyap C., "Block-based static timing analysis with uncertainty," in *Proc. of ICCAD*, 2003, 9-13, pp. 604-614.

[4] Patil D. *et al.*, "A new method for design of robust digital circuits," in *Proc. of ISQED*, 2005, pp. 676-681.

[5] Singh J. *et al.*, "Robust gate sizing by geometric programming," in *Proc. of DAC*, 2005, pp. 315-320.

[6] Srivastava A. *et al.*, "Statistical optimization of leakage power considering process variations using dual-Vth and sizing," in *Proc. of DAC*, 2004, pp. 773 – 778.

[7] Tschanz J. *et al.*, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *ISSCC Tech. Dig.*, pp. 422-423, 2002.

[8] Narendra S. *et al.*, "Impact of using adaptive body bias to compensate die-to-die Vt variation on within-die Vt variation", in *Proc. of ISLPED*, 1999, pp. 229-232.

[9] Keshavarzi A. *et al.*, "Effectiveness of reverse body bias for leakage control in scaled dual Vt CMOS ICs," in *Proc. of ISLPED*, 2001, pp. 207-212.

[10] Martin S. *et al.*, "Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads," in *Proc. of ICCAD*, 2002, pp. 721-725.

[11] Chen T. *et al.*, "Comparison of Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV) for improving delay and leakage under the presence of process variation," *IEEE Trans. VLSI on Systems*, v. 11, no. 5, 2003, pp. 888-899.

[12] Wang H. *et al.*, "Variable tapered pareto buffer design and implementation allowing run-time configuration for low-power embedded SRAMs," *IEEE Trans. on VLSI*, Vol. 13, No. 10, Oct. 2005.

[13] Prekopa A., Stochastic Programming, *Kluwer Academic*, 1995.

[14] Boyd S., Vandenberghe L., *Convex Optimization*, Cambridge, 2004.

[15] Ben-Tal, A., Nemirovski, A. "Robust solutions to uncertain linear programs," *OR Letters*, 25, 1–13, 1999.

[16] Ben-Tal A. *et al.*, Adjustable robust solutions of uncertain linear programs, *Mathematical Programming*, Volume 99, Issue 2, Mar 2004, pp. 351 – 376.

[17] Borkar S. *et al.*, "Parameter variation and impact on circuits and microarchitecture," in *Proc. of DAC*, 2003, pp. 338-342.

[18] Taur Y. and Ning H. T., Fundamentals of Modern VLSI Devices, Cambridge University Press, 1998.

[19] Berkelaar M. and Jess J., "Gate sizing in MOS digital circuits with linear programming," in *Proc. of DATE*, 1990, pp. 217-221.

[20] Rao R. *et al.*, "Statistical analysis of subthreshold leakage current for VLSI circuits," *IEEE Trans. on VLSI Systems*, 12(2), pp. 131-139, February 2004.

[21] Srivastava A. *et al.*, "Accurate and efficient parametric yield estimation considering correlated variations in leakage power and performance," in *Proc. of DAC*, 2005, pp.535-540.

[22] Azizi N. and Najm F. N. "Compensation for within-die variations in dynamic logic by using body-bias," in *Proc. of NEWCAS*, 2005, pp. 167-170.

[23] Papoulis A., *Probability Random Variables and Stochastic Processes*, New York: McGraw-Hill, 1984.

[24] Ben-Tal, A. Nemirovski, A., "Robust Convex Optimization," *Math. Oper. Res.*, 23, 1998.

[25] Mani M. and Orshansky M., "Application of fast SOCP based statistical sizing in the microprocessor design flow," in *Proc. of GLS-VLSI*, 2006, pp. 372-375.

[26] Alizadeh F. and Goldfarb D., "Second-order cone programming", Technical Report RRR, Report number 51-2001, RUTCOR, Rutgers University.

[27] http://www.mosek.com/documentation.html#manuals.

[28] Cao Y. *et al.*, "New paradigm of predictive MOSFET and interconnect modeling for early circuit design," *Proc. of IEEE CICC*, 2000, pp. 201-204.