Evaluating Statistical Power Optimization

Jason Cong, *Fellow, IEEE,* Puneet Gupta, *Member, IEEE,* and John Lee, *Student Member, IEEE*

***Abstract*—In response to the increasing variations in integrated-circuit manufacturing, the current trend is to create designs that take these variations into account statistically. In this paper, we quantify the difference between the statistical and deterministic optima of leakage power while making no assumptions about the delay model. We develop a framework for deriving a theoretical upper bound on the suboptimality that is incurred by using the deterministic optimum as an approximation for the statistical optimum. We show that for the mean power measure, the deterministic optima is an excellent approximation, and for the mean plus standard deviation measures, the optimality gap increases as the amount of inter-die variation grows, for a suite of benchmark circuits in a 45 nm technology. For large variations, we show that there are excellent linear approximations that can be used to approximate the effects of variation. Therefore, the need to develop special statistical power optimization algorithms is questionable.**

***Index Terms*—Algorithms, gate sizing, optimization, physical design, statistical power.**

I. Introduction

STATISTICAL optimization via circuit sizing has been an active research topic over the last decade. The realization was that the traditional corner-based optimization [2]) may be too pessimistic [3], and the trend was to incorporate more and more statistical data into the optimization process.

There are many papers that explore the benefits of adding statistical delay data into the optimization process [4], [5]–[10], and there are also a number of papers that use a statistical power measure [7], [11]–[15]. However, to the best of our knowledge, there is no publication that shows the benefits of using the statistical power measure alone.

This brings up an interesting question: how much of the improvement should be attributed to the use of a *statistical* delay model and how much should be attributed to the use of a *statistical* power model? This question is part of a growing skepticism over the benefits of statistical optimization, and whether they outweigh its costs. Adopting statistical analyses and optimization involve considerable overhead in terms of the engineering effort involved as well as the turn-around time. It requires an almost complete overhaul of process modeling, circuit simulation, and also a modification to the algorithms for statistical optimization. It is, therefore, important to do a thorough cost-benefit analysis of statistical optimization compared to conventional deterministic optimization methods.

The related question of statistical delay optimization vs. deterministic delay optimization has already been studied [16]–[18]. In [16], the claim is made that corner or scenario-based optimization is still the most practical because of the following:

1) intra-die effects are still small;
2) there is usually not enough information to do a full-blown statistical analysis;
3) the gains of using a full-blown statistical analysis are small.

In [18], the authors quantify the difference between corner-based methodologies and full statistical optimization methods. They find that with a 5% variation in stage delay, the full-blown statistical analysis and optimization gives a mere 2% improvement, and a 12% variation gives a 6% improvement over a statistical worst-case corner that employs a guardband. In [17], the tradeoff between yield and circuit delay, and the improvements in slack are examined. Significant improvements are shown for a set of benchmark circuits.

In this paper, we focus on the amount of improvement that can be made by using a statistical power measure as an objective for gate sizing, l_{eff} and V_{th} assignment, when compared to the deterministic power measure. The key contributions of the paper are as follows.

1) We develop a mathematical programming-based framework to estimate the suboptimality gap between different power measures.
2) For the common case of discrete gate sizing, we give an intuitive explanation of the suboptimality using solution rank orders.
3) We show that for certain sizes of variations, the deterministic power measure is a provably good approximation for the statistical power measures, which means that the deterministic power measure can be used in place of the statistical power measures with very similar optimization results.
4) We present a simplified measure of statistical power that can be used as a proxy for full statistical power optimization.

It is important to mention that the analysis in this paper is independent of the model for the delay. This paper does not give judgments on the difference between statistical and

Manuscript received July 6, 2009; revised January 20, 2010 and May 3, 2010; accepted June 4, 2010. Date of current version October 20, 2010. This work was supported in part by the SRC, under Contract 2006-TJ-1460, in part by the NSF Award 811832, and in part by the UC Discovery Grant ele07-10283 under the IMPACT Program. A preliminary version of this work appeared in [1]. This paper was recommended by Associate Editor M. Orshansky.

J. Cong is with the Department of Computer Science, University of California, Los Angeles, CA 90095 USA (e-mail: cong@cs.ucla.edu).

P. Gupta and J. Lee are with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: puneet@ee.ucla.edu; lee@ee.ucla.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCAD.2010.2061390

0278-0070/$26.00 © 2010 IEEE

TABLE I
NOTATIONS

Symbol	Meaning
$w_i, \vec{w}$	Width of gate i/vector of gate widths
$l_i, \vec{l}$	Length of gate i/vector of gate lengths
$v_{\mathrm{th}_i}, \vec{v}_{\mathrm{th}}$	Threshold voltage of gate i/vector of threshold voltages
$\vec{z}$	Adjusted gate widths $z_i = w_i e^{\alpha l_i^2 + \beta l_i} e^{-\gamma v_{\mathrm{th}_i}}$
$\mathbf{L}_{\mathrm{w}}$	Within-die gate length variation random variable
$\mathbf{L}_{\mathrm{d}}$	Die-to-die gate length variation random variable
$\mathbf{V}_{\mathrm{th_w}}$	Within-die v_{th} variation random variable
$\mathbf{V}_{\mathrm{th_d}}$	Die-to-die v_{th} variation random variable

TABLE II
MEASURES OF STATISTICAL LEAKAGE POWER

Measure	Symbol	Expression
Deterministic	$p_d(\cdot)$	$\sum \kappa_i$
Mean	$p_m(\cdot)$	$\sum \kappa_i \exp((\lambda^2_{\mathrm{w}(i)}\sigma^2_{\mathbf{L}_{\mathrm{w}(i)}} + \lambda^2_{\mathrm{d}(i)}\sigma^2_{\mathbf{L}_{\mathrm{d}}} + \ldots$ $\tau^2_{\mathrm{w}(i)}\sigma^2_{\mathbf{V}_{\mathrm{th_w}(i)}} + \tau^2_{\mathrm{d}(i)}\sigma^2_{\mathbf{V}_{\mathrm{th_d}}})/2)$
Mean + 3σ	$p_{m3\sigma}(\cdot)$	$p_m(\cdot) + 3p_\sigma(\cdot)$
99.87% – quantile[1]	$p_q(\cdot)$	$\sum \kappa_i \exp(3(\lambda_{\mathrm{d}(i)}\sigma_{\mathbf{L}_{\mathrm{d}}} + \tau_{\mathrm{d}(i)}\sigma_{\mathbf{V}_{\mathrm{th_d}}}))$
Standard deviation, σ	$p_\sigma(\cdot)$	[see (4)]

[1]For inter-die variation only.

deterministic delay optimization, or the difference between static and statistical static timing analysis. The delay is only used to generate examples for the suboptimality bounds.

The approach in this paper intends to remove the dependence on timing feasibility, which is a consideration that makes the problem of finding bounds difficult. To remove this dependence, we rely on the structure of the power objective to create a relaxation of the timing feasible region. This can be done because a power optimum inherently contains information about the timing feasible region, as this optimum is the minimum power point in the region. Once this is done, the relaxation can be used to find bounds without any further dependence on the timing feasible region.

The rest of the paper is organized as follows. The following section outlines the leakage power measures and models used in this paper. Section III describes the test circuits that are used in this paper. Section IV gives an intuitive explanation of the suboptimality gap using rank orders. Section V develops the mathematical framework to estimate the suboptimality gap incurred from using the deterministic power measure in place of the statistical power measures. Section VI introduces a simplified measure of statistical power that can be used as a proxy for full statistical power optimization. The paper is summarized in Section VII.

II. STATISTICAL POWER

In this paper, uppercase bold symbols represent random variables (e.g., $\mathbf{X}$) and uppercase non-bold symbols represent commonly used constants (e.g., V_{dd}). Vector quantities will have arrows above them (e.g., $\vec{x}$) and scalar quantities will be lowercase non-bold (e.g., p). The principal symbols are summarized in Table I.

This paper examines the benefits of using the statistical power as an objective to the problem as follows:

$$\begin{array}{ll} \text{minimize} & \text{Statistical Power}(w, l, v_{\mathrm{th}}) \\ \text{subject to} & \text{Delay}(w, l, v_{\mathrm{th}}) \leq T_{\max} \end{array} \tag{1}$$

compared to using the deterministic power as the objective. In this section, we derive expressions for statistical leakage power, review its mathematical properties, and discuss how the different properties of the statistical power would affect the resulting optima.

A. Variations

In this paper, the gate length and the threshold voltage are assumed to be the sources of power variations. These variations are assumed to be Gaussian in both the length and the threshold voltage. The standard deviations in v_{th} are 4.714% in both die-to-die (dtd) and within-die cases (wid); the within-die variations are uncorrelated and they affect each gate independently. Three different standard deviations for the length are used in this paper:

1) 1 nm (dtd) and 0.5 nm (wid);
2) 1.6 nm (dtd) and 1.15 nm (wid);
3) 2 nm (dtd) and 0 nm (wid).

These are representative of the variations for 45 nm given in the International Technology Roadmap for Semiconductors Roadmap 2007.

The effects of the variation are simulated for the gates in the Nangate Open Cell Library v1.2 [19] using predictive technology model 45 nm and HSPICE 2007.[1] We assumed that the input combinations to each gate are equi-probable and the measured leakage power is the average power over all the input combinations.

B. Models

The leakage power is modeled as a log-normal random variable as follows [20]:

$$\begin{aligned} \mathbf{P}_{\mathrm{L}} &= \sum \kappa_i e^{\lambda_{\mathrm{d}(i)}\Delta\mathbf{L}_{\mathrm{d}} + \lambda_{\mathrm{w}(i)}\Delta\mathbf{L}_{\mathrm{w}(i)}} e^{\tau_{\mathrm{d}(i)}\Delta\mathbf{V}_{\mathrm{th\,d}} + \tau_{\mathrm{w}(i)}\Delta\mathbf{V}_{\mathrm{th\,w}(i)}} \\ &= \sum \mathbf{P}_{\mathrm{L}_i} \end{aligned} \tag{2}$$

where κ_i is the nominal power for the gate i; $\Delta\mathbf{L}_{\mathrm{w}(i)}$, $\Delta\mathbf{L}_{\mathrm{d}}$, $\Delta\mathbf{V}_{\mathrm{th\,w}(i)}$, and $\Delta\mathbf{V}_{\mathrm{th\,d}}$ are independent Gaussian random variables; and $\lambda_{\mathrm{d}(i)}$, $\lambda_{\mathrm{w}(i)}$, $\tau_{\mathrm{d}(i)}$, and $\tau_{\mathrm{w}(i)}$ are coefficients that are used to fit the mean and the standard deviation of the SPICE-simulated data. In (2) above, the variations in the threshold voltage and the gate length are assumed to be Gaussian, and the relationship between these parameters to the leakage power is exponential.

To model the change in the leakage as a function of the size, threshold voltage and the length (see Section V), we use

[1]For reference, the channel length used in the HSPICE models was 50 nm. Also, note that this simulation tool is different from the one in [1]. It was changed because the version of the tool in [1] could not simulate the effects of v_{th} variation.

the approximation as follows:

$$\kappa_i \approx \kappa_i' w_i e^{\alpha_i l_i^2 + \beta_i l_i} e^{-\gamma_i v_{\mathrm{th}_i}} = \kappa_i' z_i \tag{3}$$

where κ_i', α_i, β_i, and γ_i are parameters that are fitted to the data. The variable z_i denotes the "adjusted gate width," which incorporates the effect of w, l, and v_{th} on the power into an equivalent gate width. The z variables cannot be re-mapped to a unique w, l, and v_{th}, but this is not a problem, as z is used only to compute lower bounds and is not used for design purposes. The nonlinear mapping from w, l, and v_{th} to z has the effect of shifting the nonlinear relation of p_d on v_{th} to a linear one in z (which is, in turn, nonlinear in v_{th}), creating a useful abstraction for computing bounds.

C. Measures of Statistical Leakage Power

In this paper, we will cover the statistical measures summarized in Table II: the deterministic, mean, mean + 3σ, and the quantile power measures. The mean + 3σ measure refers to the mean + 3σ of the total leakage power of a design. When there is no within-die variation, the 99.87% quantile measure, which corresponds to the three sigma quantile of a Gaussian distribution, will be used as well.

In the table, $p_\sigma(\cdot)$ is the measure of the standard deviation of the power, which can be expressed as follows:

$$p_\sigma^2 = \sum_i \sum_j \mathbf{E}\left[\mathbf{P}_{\mathrm{L}_i}\mathbf{P}_{\mathrm{L}_j}\right] - \mathbf{E}\left[\mathbf{P}_{\mathrm{L}_i}\right]\mathbf{E}\left[\mathbf{P}_{\mathrm{L}_j}\right]. \tag{4}$$

The quantile measure p_q is used only when inter-die variation is present. When intra-die, or within-die, variation is present, there is no closed-form expression for p_q.

The power measures above have useful mathematical properties. p_d, p_m, and p_q are all linear in κ_i, and are thus concave *and* convex in z_i. $p_{m3\sigma}$ is convex in $\vec{z}$.

D. Why Do We Expect the Optimizations to be Similar?

The statistical power and deterministic power are not similar. For example, the statistical leakage power can be larger than the deterministic leakage by 10–500%. It is natural to expect that the influence of these measures will also be different, and that optimizing statistical power will yield different results compared to deterministic power.

In optimization, however, it is not the magnitude of the power but the *relative* magnitude which matters. If the statistical power is a scaled version of the deterministic power, then the optima will be the same. To see this mathematically, we examine the optimality condition for an optimum $x^\star$: $x^\star$ is optimal if for any *feasible* $x^\star + \Delta x$

$$f(x^\star) \leq f(x^\star + \Delta x). \tag{5}$$

This condition will also hold for any *positive* scaling of $f(x)$.

Although the values of the statistical power and the deterministic powers may be quite different, the trends are similar; the mean power is larger than the deterministic power for all of the gates, as is the quantile power, and so on. For example, consider Fig. 1, which shows the power vs. size sensitivities. The sensitivities for the different gates follow the same trend. This suggests that the optimizations will be similar as well.

To see why this happens for statistical power, we take the quantile power expression with inter-die length variation as an

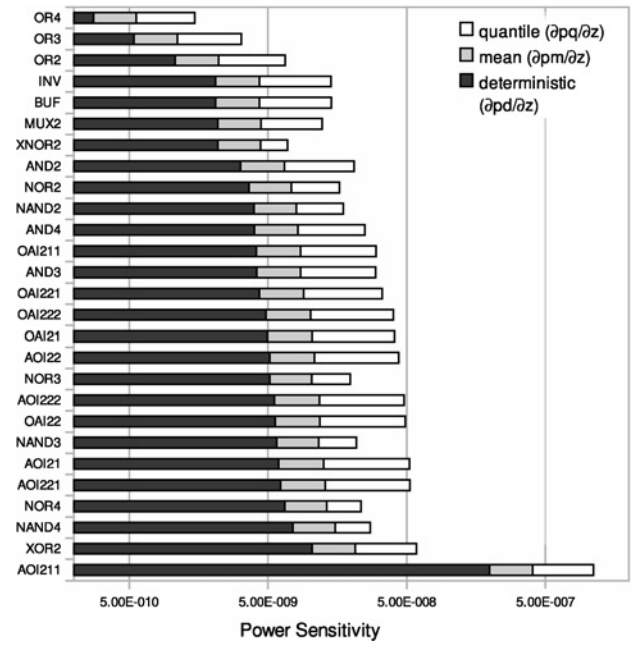


Fig. 1. Power sensitivities for the different gates in the Nangate Library. The power vs. size sensitivities of the deterministic power measure ($\partial p_d/\partial z$), the mean power measure ($\partial p_m/\partial z$), and the quantile measure ($\partial p_q/\partial z$) for σ_L = 2 nm (dtd)/0 nm (wid) and $\sigma_{V_{\mathrm{th}}}$ = 4.7% (dtd)/0% (wid) are shown. The sensitivities are sorted by the p_d. Notice that a sorting by p_m would be very similar, while a sort by p_q would be significantly different.

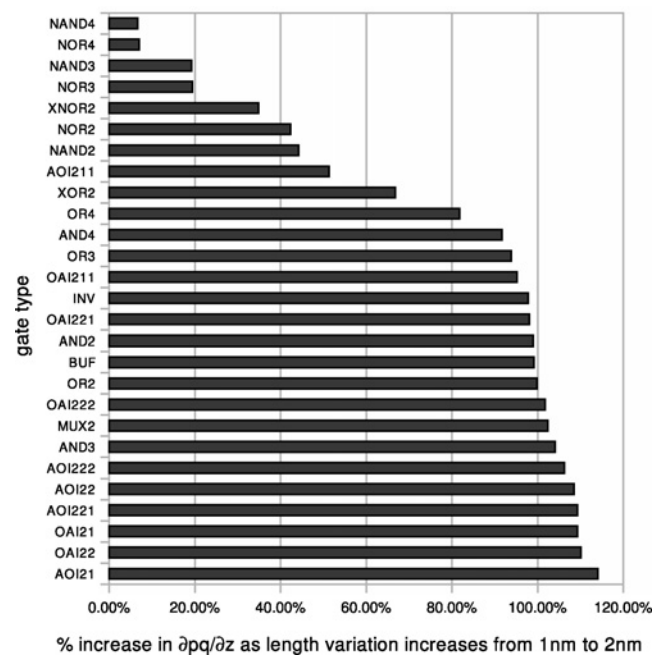


Fig. 2. Percentage increase in the quantile power vs. size sensitivity ($\partial p_q/\partial z$) as σ_{L} = 1 nm increases to 2 nm is shown. The increase in the sensitivity is not seen equally for all the gates.

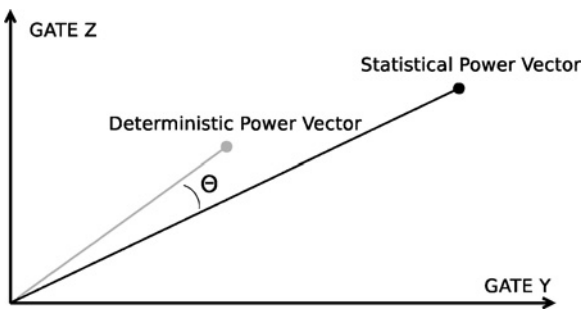


Fig. 3. Visual representation of the power vector for deterministic power and statistical power. Each axis represents the power for a different gate type. The vectors have different magnitude and direction, but the angle Θ between them is very small.

example

$$p_q = \sum \kappa_i e^{3\lambda_{\mathrm{d}(i)}\sigma_{\Delta \mathbf{L}_\mathrm{d}}}. \tag{6}$$

If the $\lambda_{\mathrm{d}(i)}$ are all equal, then the effect of variation will be seen equally for each gate and the objective will be a scaled version of the deterministic power as follows:

$$\begin{aligned} p_q &= \sum \kappa_i e^{3\lambda_{\mathrm{d}(i)}\sigma_{\Delta \mathbf{L}_\mathrm{d}}} \\ &= e^{3\lambda_\mathrm{d}\sigma_{\Delta \mathbf{L}_\mathrm{d}}} \sum \kappa_i \\ &= e^{3\lambda_\mathrm{d}\sigma_{\Delta \mathbf{L}_\mathrm{d}}} p_d. \end{aligned} \tag{7}$$

In actuality, the values of $e^{3\lambda_{\mathrm{d}(i)}\sigma_{\Delta \mathbf{L}_\mathrm{dtd}}}$ are different for different gates. Fig. 2 plots the percentage increase in the power vs. size sensitivities, $\partial p_q/\partial z$, from $\sigma_\mathrm{L} = 1$ nm to 2 nm. The percentage increase is not uniform across the different gates, indicating that the $\lambda_{\mathrm{d}(i)}$ are not equal. The larger variations thus magnify the discrepancies between the different gates.

Surprisingly, the effect of the scaling leaves the orientation of the power sensitivities mostly intact. The magnitude changes significantly, but the direction of the vector remains similar. We can plot a power vector where each gate is a separate dimension, and the magnitude along the dimension is the leakage power for that gate. We can compute the difference in the angle of the corresponding statistical power vector and the deterministic power vector, θ, as shown in Fig. 3. For $\sigma_{\Delta \mathbf{L}} = 2$ nm (dtd)/0 nm (wid) and $\sigma_{\Delta \mathbf{V}_\mathrm{th}} = 4.7\%$ (dtd and wid), the difference between the vectors is very small, at $\Theta \approx 1$ deg. In Section V, we will see that the effects may not be large enough to make a significant difference in the optimized powers.

E. Increasing Variations

It is also interesting to examine how the sensitivities change as the variations increase. We can use (2) for a first-order analysis of how the sensitivities will change as the variations change.

We reconsider the case with inter-die length variation only. Doubling the variations results in the following:

$$\begin{aligned} p_q &= \sum \kappa_i e^{3\lambda_{\mathrm{d}(i)}(2\sigma_{\Delta \mathbf{L}_\mathrm{d}})} \\ &= \sum (e^{3\lambda_{\mathrm{d}(i)}\sigma_{\Delta \mathbf{L}_\mathrm{d}}})^2 \kappa_i. \end{aligned} \tag{8}$$

TABLE III
NUMBER OF GATES IN TEST CIRCUITS AND THE CORRESPONDING DELAY TARGET IN PARENTHESIS (IN PS)

	v1	v2	v3	v4
c432	118 (431)	137 (494)	93 (557)	81 (619)
c499	586 (477)	322 (612)	189 (746)	174 (881)
c880	346 (408)	237 (562)	212 (716)	176 (871)
c1355	598 (473)	356 (599)	222 (724)	174 (850)
c1908	676 (802)	571 (863)	533 (923)	528 (990)
c2670	1041 (610)	791 (787)	748 (964)	723 (1140)
c3540	1615 (1096)	1359 (1289)	1107 (1482)	1046 (1675)
c5315	2019 (1090)	1867 (1226)	1766 (1362)	1717 (1498)
c6288	3089 (1845)	1656 (4127)	1780 (6410)	1189 (8692)
c7552	2894 (866)	2640 (977)	2481 (1089)	2425 (1200)
alu	12 598 (896)	5190 (8460)	5271 (16 025)	3237 (23 652)

TABLE IV
GATE SIZES IN THE NANGATE LIBRARY

Gates	Sizes
HA, FA, OAI33	1x
MUX2, XNOR2, XOR2	1x–2x
AND2/3/4, AOI211/21/221/222/22, NAND2/3/4, NOR2/3/4, OAI211/21/221/222/22, OR2/3/4	1x–4x
BUF, INV	1x–32x

This will increase the difference in the scaling factors between gates. This is seen in Fig. 2, where the increase in the variations changes the sensitivities for each gate. Notice that the sensitivities do not increase uniformly for each gate. The growth of the sensitivities depends on the topology of the design, and how much effect the variations will have on the gate.

III. CIRCUIT EXAMPLES

In this paper, we use the International Symposium on Circuits and Systems '85 benchmarks and a 128-bit arithmetic logic unit (ALU) [21] as examples. The Verilog register-transfer level (RTL) of these benchmarks are synthesized to four different target speeds using the Encounter RTL compiler [22] to the the Nangate Open Cell Library v1.2 [19], which uses the 45 nm technology node and has the sizes given in Table IV. The synthesized speeds are the maximum speed, the minimum speed, and two speeds in between. The fastest speed is labeled v1 and the slower synthesized speeds have higher cardinality (e.g., the slowest speed is v4). A table listing the number of gates in each design and the target delays is given in Table III. The tighter delay constraints result in larger designs, as these designs utilize more buffering to meet the delay constraint. These larger synthesized designs are also more interesting from a sizing point of view, as the space of possible solutions is larger.

All the optimization routines in this paper are solved using the MATLAB Optimization Toolbox [23].

IV. COMPARING RANDOMLY GENERATED CONFIGURATIONS

The major difficultly in estimating the difference between statistical and deterministic optimization is due to the difficulty

TABLE V
RANK CORRELATION τ

	p_m				$p_{m3\sigma}$			
	v1	v2	v3	v4	v1	v2	v3	v4
c432	0.989	0.986	0.985	0.985	0.653	0.591	0.603	0.604
c499	0.990	0.988	0.990	0.990	0.685	0.605	0.516	0.524
c880	0.989	0.989	0.989	0.987	0.658	0.651	0.649	0.585
c1355	0.995	0.987	0.987	0.991	0.684	0.603	0.504	0.537
c1908	0.988	0.987	0.987	0.987	0.671	0.624	0.608	0.606
c2670	0.991	0.991	0.991	0.992	0.712	0.647	0.646	0.651
c3540	0.991	0.990	0.989	0.989	0.702	0.668	0.619	0.602
c5315	0.991	0.991	0.991	0.991	0.675	0.659	0.641	0.634
c6288	0.989	0.999	0.999	0.999	0.680	0.894	0.913	0.922
c7552	0.990	0.990	0.990	0.990	0.686	0.660	0.637	0.642
ALU	0.997	0.999	0.999	0.999	0.894	0.899	0.885	0.905

in gate sizing itself. With discrete sizes (e.g., $\vec{w} \in \{1, 2, 4, 8\}^n$ and $\vec{l} \in \{1, 2, 3\}^n$), the problem is NP-complete [24] and it is tremendously difficult to find an optimal solution. However, there is some intuition that can be gained by examining the statistical and deterministic powers of points even if timing feasibility is ignored. Namely, we can consider: 1) how does a sorting of configurations (w, l, and v_{th} assignments) change when ordered statistically, instead of deterministically, and 2) how does the statistical power vary for configurations with the same deterministic power?

As an experiment, the deterministic and statistical powers were computed for 10 000 randomly generated widths, lengths, and v_{th}, for each design. Three different v_{th} values were used (low, high, and normal), and four different gate lengths were used ({+0 nm, +1 nm, +2 nm, +3 nm}). In these examples, we assumed that $\sigma_L = 2\,\text{nm(dtd)}/0\,\text{nm(wid)}$, and $\sigma_{V_{\text{th}}} = 4.7\%$ (dtd & wid).

Fig. 4 plots the deterministic power vs. the statistical power for the c1355 v2 circuit. From the plot, it is clear that the p_d and the p_m measures correlate more than the p_d and the $p_{m3\sigma}$. We can measure the correlation between the ranks using Kendall's τ [25]. Sequences that are perfectly correlated (e.g., have the same ranking) have a $\tau = 1$ and sequences that are perfectly anti-correlated (e.g., are reverses of each other) have a $\tau = -1$. The values for the different designs are given in Table V.

From this table, we can see that the p_m rankings are near perfect (≈ 1). This indicates that the optimizations will be very similar. However, the $p_{m3\sigma}$ rankings range from 0.5 to 0.9. In this case, it is not clear from the rankings whether the statistical power is a good surrogate for the deterministic power.

Why the p_m rankings correlate better than the $p_{m3\sigma}$ measures can be seen in Fig. 1, which plots the power vs. size sensitivities for the p_d, p_m, and p_q measures (the $p_{m3\sigma}$ can be thought of as a rough approximation of the p_q measure). Notice that a sorting of the gates by sensitivity would be very similar for the p_d and the p_m measures. However, the sorting by p_q, and, hence, the relations of the sensitivities of different gates, is very different.

In the analysis above, it is difficult to tell exactly how large the gap will be between the deterministic and statistical sizing solutions. To get a direct idea of the difference between the deterministic power compared to the statistical power,

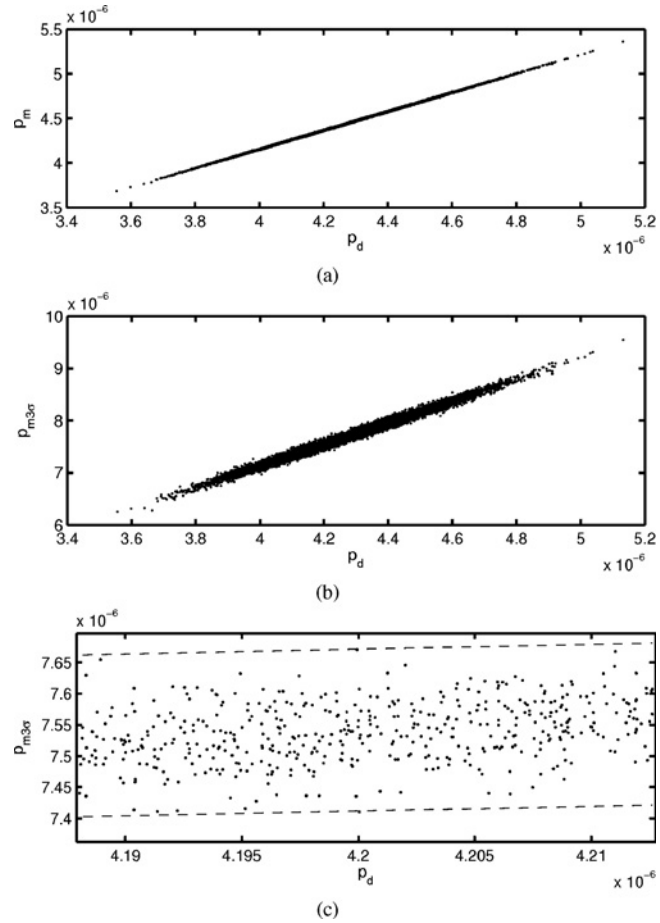


Fig. 4. Plot of the deterministic power (x-axis) vs. statistical power (y-axis) for the c1355 v2 circuit for width sizing only. (a) Plots for the p_d vs. the p_m measure, whose relation is nearly linear. In contrast, the p_d vs. $p_{m3\sigma}$ measure in (b) has more noise. If the optimal deterministic power ($p_d^\star$) is known, we can focus on a region of the plot where $p_d \approx p_d^\star$ to estimate the suboptimality at this power value. (c) Region of the p_d vs. $p_{m3\sigma}$ plot is chosen around $p_d = 4.2 \cdot 10^{-6}$. Upper bounding and lower bounding lines are added to the figure. The difference in $p_{m3\sigma}$ between the two lines (approximately $2.6 \cdot 10^{-7}$ or about 3.5%) is an estimate of the suboptimality bound. Compare this with the bounds in Table VI.

we can make a useful assumption. Suppose we know the value of the minimum deterministic power ($p_d^\star$) that is timing feasible. Then, we can estimate the difference between the best statistical sizing and the worst sizing for the given deterministic power as in Fig. 4(c).

There is inherent error in this process because it depends on the number of samples. As the number of samples grows, the gap will increase as the number of extreme points are sampled. In the following section, we will compute these bounds without needing to sample the distribution.

V. SUBOPTIMALITY BOUNDS

The central question in this paper is whether the deterministic power solution is a good approximation for the statistical power optimum. This generally requires information about the space of timing-feasible solutions ($\mathcal{T}$), which is difficult to describe, and is highly problem dependent, making it hard to give an exact answer. However, there is a way to solve a simpler problem with no assumption on the structure of $\mathcal{T}$,

instead, relying on the structure of the deterministic power objective.

In this section, we consider the following question. Suppose we approximate the solution to the statistical power optimization problem as follows:

$$\begin{array}{lll} \text{minimize} & p_s(\vec{w}) & \text{(statistical power)} \\ \text{subject to} & \vec{w} \in \mathcal{T} & \text{(timing constraint} \\ & & \text{\& discreteness)} \\ & \vec{w} \in \mathcal{B} & \text{(upper and lower)} \\ & & \text{(size bounds on } \vec{w}) \end{array} \tag{S}$$

using the deterministic power optimum, $\vec{w}_d^\star$, which is the solution to the problem as follows:

$$\begin{array}{lll} \text{minimize} & p_d(\vec{w}) & \text{(deterministic power)} \\ \text{subject to} & \vec{w} \in \mathcal{T} & \text{(timing constraint} \\ & & \text{\& discreteness)} \\ & \vec{w} \in \mathcal{B} & \text{(upper and lower).} \\ & & \text{(size bounds on } \vec{w}). \end{array} \tag{D}$$

How good of an approximation will this be, and what is a bound for the suboptimality of this solution?

In the following, we will describe a method for creating suboptimality bounds. First, a simple set $\mathcal{T}'$ is constructed that contains $\mathcal{T}$. Optimizing the statistical power over this simpler set will return a lower bound on the statistical power optimum. This lower bound is then compared with the statistical power of the approximate solution, $p_s(\vec{w}_d^\star)$, to bound the accuracy of the approximation. This is described in detail below.

A. Relaxed Constraints, Enclosing Sets, and Lower Bounds

The difficult part of w, l, and v_{th} optimization is the timing constraint and the discreteness constraint. Thus, to find a quick lower bound, we must first relax the timing constraint with a looser constraint. In other words, we would like to *relax* the constraints, by enclosing the timing feasibility and discreteness condition in a simple, convex set.

Relaxing the constraints of a problem turns the resulting solution into a lower bound for the true solution. For example, consider the sets $\mathcal{T}_0 \subseteq \mathcal{T}_1 \subseteq \ldots \subseteq \mathcal{T}_k$ and the following sequence of problems:

$$\begin{array}{llll} (\mathrm{P}_i) & \text{minimize} & p_s(\vec{w}) \\ & \text{subject to} & \vec{w} \in \mathcal{T}_i \\ & & \vec{w} \in \mathcal{B} \subseteq \mathbb{R}^n. \end{array} \tag{9}$$

If the optimal solution of problem (P_i) is $\vec{w}_i^\star$, then we have the property as follows:

$$p_s(\vec{w}_0^\star) \geq p_s(\vec{w}_1^\star) \geq \ldots \geq p_s(\vec{w}_k^\star). \tag{10}$$

In other words, the optimal value for the relaxed problem is a lower bound for the original problem.

The intuition for this is the fact that the constraints in the relaxed problem enclose the constraints on the original problem. Thus, the optimal solution in the original problem is also *feasible* for the relaxed problem. In the process of solving the relaxed problem, the solver is free to choose a better point in the larger space, making the resulting optimum a *lower bound* for the original problem.

B. Linear Functions, Optimum Solutions, and Enclosing Sets

For certain classes of functions, it is easy to find a simple set that encloses the optimum. The following analysis will derive a set using the properties of linear functions, but the results also hold for more general functions.[2]

The key to finding an enclosing set for the constraints is to start with an optimal solution and leverage the fact that any other feasible point cannot be better. For example, if $\vec{w}_d^\star$ is optimal for (D), then

$$\forall \vec{w} \in (\mathcal{T} \cap \mathcal{B}): \;\; p_d(\vec{w}_d^\star) \leq p_d(\vec{w}). \tag{12}$$

For linear functions, the inequality on the right side can be rewritten in a simple form. This is because any linear function $f(x)$ can be expressed in the form $f(x) = f(x_0) + s^T(x - x_0)$, where $s = \nabla f(x)$.[3] Thus, expanding about the minimum $x^\star$ of f gives the following:

$$f(x^\star) \leq f(x^\star) + s^T(x - x^\star) \tag{13}$$

$$0 \leq s^T(x - x^\star). \tag{14}$$

Applying this to (12) with $s = \nabla p_d(\vec{w}_d^\star)$ shows the following:

$$(\mathcal{T} \cap \mathcal{B}) \subseteq \mathcal{T}' = \{\vec{w} \mid 0 \leq s^T(\vec{w} - \vec{w}_d^\star)\}. \tag{15}$$

Note that $\mathcal{T}'$ is a continuous, connected set. This gives the following relaxed problem:

$$\begin{array}{ll} \text{minimize} & p_d(\vec{w}) \\ \text{subject to} & \vec{w} \in \mathcal{T}' \\ & \vec{w} \in \mathcal{B} \subseteq \mathbb{R}^n. \end{array} \tag{16}$$

C. Creating Lower Bounds for Related Problems

The above analysis seems a little circular; the optimum is required to create a lower bound for the optimum. However, the utility emerges when we use the same enclosing sets to find lower bounds for related problems.

Suppose the solution for (D) is known, and we would now like to find the lower bound for (S), which has a different objective function, but identical constraints. This can be done by leveraging the solution $\vec{w}_d^\star$ for (D) to compute a simple,

[2]Specifically, (15) holds whenever $p_d(\vec{w})$ satisfies the following:

$$\{\vec{w} \mid p_d(\vec{w}^\star) \leq p_d(\vec{w})\} \subseteq \{\vec{w} \mid 0 \leq \nabla p_d(\vec{w}^\star)^T(\vec{w} - \vec{w}^\star)\}. \tag{11}$$

Mathematically, this is equivalent to saying that $0 \leq \nabla p_d(\vec{w}^\star)^T(\vec{w} - \vec{w}^\star)$ defines a supporting hyperplane for the super-level sets $p_d(\vec{w}^\star) \leq p_d(\vec{w})$. This includes functions that are linear, concave, and quasi-concave. This does not necessarily hold for convex functions or posynomial functions, and in these cases, other properties of the timing-feasible region must be assumed for the results in this section to hold.

[3]Note that s is constant over x for linear functions.

enclosing set for the constraints, as in the section above. The relaxed problem is then solved as follows:

$$\begin{array}{ll} \text{minimize} & p_s(\vec{w}) \\ \text{subject to} & \vec{w} \in \mathcal{T}' \\ & \vec{w} \in \mathcal{B} \subseteq \mathbb{R}^n. \end{array} \tag{17}$$

The problem is solved over the continuous set $\mathcal{T}'$, and the continuous upper and lower bound constraints in $\mathcal{B}$. The corresponding solution $\vec{w}'$ can be used as a lower bound on the true optimum $w_s^\star$ ($p_s(\vec{w}') \leq p_s(\vec{w}_s^\star)$). In the case of $p_s = p_m$ and $p_s = p_q$, this is a linear programming problem, and for $p_s = p_{m3\sigma}$, this is a nonlinear convex optimization problem.

Using this lower bound, we can now find a bound for how well the deterministic solution approximates the solution for the statistical problem. The suboptimality gap between this approximation $\vec{w}_d^\star$, and the true optimum, $\vec{w}_s^\star$ is bounded by the following:

$$\delta_{\text{so}} = 100 \cdot \frac{p_s(\vec{w}_d^\star) - p_s(\vec{w}')}{p_s(\vec{w}_d^\star).} \tag{18}$$

Smaller values indicate that $\vec{w}_d^\star$ is a good approximate solution for $\vec{w}_s^\star$, and larger values indicate that it is a bad approximation. For example, if

$$\delta_{\text{so}} \leq 5\% \tag{19}$$

then $\vec{w}_d^\star$ is a 5% approximate solution for $\vec{w}_s^\star$. In other words, using $\vec{w}_d^\star$ in place of the real optimum $\vec{w}_s^\star$, would cost at most 5% (it is *suboptimal* by at most 5%).

This process also works with optimization over w, l, and v_{th}. This is exactly similar to the above example, with w replaced by z, the adjusted gate widths.[4]

A surprising fact is that no properties of $p_s(\vec{w})$ are assumed. It may be non-convex or it may be nonlinear, as in the case of the $p_{\text{m}3\sigma}$ measure. The only assumption is that (17) is solvable.

Another interesting fact is that the deterministic optima is only used to determine: 1) the minimum deterministic power that is timing feasible, and 2) its corresponding statistical power. The delay of the design is thus important in determining the deterministic optimum, but does not play any other role in the bounding process. This highlights the independence of this method on the timing feasible region.

[4]This gives the following problem:

$$\begin{array}{ll} \text{minimize} & p_s(\vec{z}) \\ \text{subject to} & 0 \leq \nabla p_d(\vec{z}_d^\star)^T(\vec{z} - \vec{z}_d^\star) \\ & z_{min} \leq \vec{z} \leq z_{max}. \end{array} \tag{20}$$

Here, z is a continuous variable, which represents the effect of the widths, lengths, and v_{th} on the power. $z_{\min}$ and $z_{\max}$ are the minimum and maximum values of $\vec{z}$. For example, for the ith entry, they are

$z_{i,\min} = w_{i,\min} e^{\alpha l_{i,\max}^2 + \beta l_{i,\max}} e^{-\gamma v_{\text{th}_i,\max}}$ and

$z_{i,\max} = w_{i,\max} e^{\alpha l_{i,\min}^2 + \beta l_{i,\min}} e^{-\gamma v_{\text{th}_i,\min}}$.

Interestingly, the actual values of $\vec{l}$, $\vec{w}$, and v_{th} do not play a direct role in the optimization above. They affect the optimization by determining a range for the values of z_i. In fact, the corresponding values of w_i, l_i, and v_{th} may not be unique; it is only important that there is at least one combination of w_i, l_i, and v_{th} that satisfies $z_i = w_i e^{\alpha l_i^2 + \beta l_i} e^{-\gamma v_{\text{th}_i}}$. Thus, z acts as a proxy for w, l, and v_{th} and a corresponding value is not important, as z is used solely to find a lower bound on the power, and not a minimum power configuration.

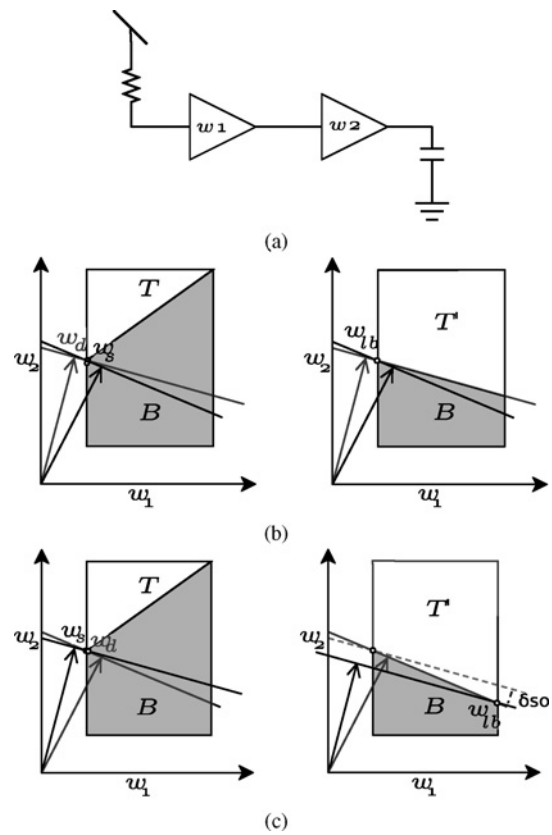


Fig. 5. Suboptimality examples for the (a) two gates. The examples in (b) and (c) have the same optimal solutions for both the statistical and deterministic measures, which means that the actual suboptimality is zero. However, the two examples have different suboptimality *bounds*.

A visual example of the lower bounding process is shown in Fig. 5. The figure shows that the suboptimality bound is more related to the geometry of the problem than the actual difference between the statistical and deterministic optima.

D. Width-Sizing Experiment

In this section, we compute suboptimality bounds for width sizing. The bounds (δ_{so}) are computed for each circuit in Section III and the results are presented in Table VI.

The Nangate Library is modeled using (3). The synthesized widths from the Cadence RTL compiler are assumed to be the optimal deterministic widths,[5] $\vec{w}_d^\star$.

These tables show that the bounds grow larger as the proportion of die-to-die variation increases. Note that this is in spite of the fact that the total σ_L per gate is roughly the same, for example

$$\sigma_{L,\text{total}}^2 \approx 1.15^2 + 1.6^2 \approx 2^2. \tag{21}$$

This is because the die-to-die variations effect the power more efficiently than within-die variations, which may cancel each other out [see (4)]. The case $\sigma_{\text{L}} = 1$ nm (dtd) /0.5 nm (wid) is given for comparison with [1].

[5]We will drop the assumption that the synthesized widths are optimal in Section V-E.

TABLE VI

SUBOPTIMALITY (δ_{so}) FOR LEAKAGE OPTIMIZATION WITH l AND V_{th} VARIATIONS

	p_m					$p_{m3\sigma}$				
	v1 (%)	v2 (%)	v3 (%)	v4 (%)	Avg	v1 (%)	v2 (%)	v3 (%)	v4 (%)	Avg
σ_L = 1 nm (dtd)/0.5 nm (wid) and $\sigma_{V_{th}}$ = 4.7% (dtd and wid)										
c432	0.34	0.31	0.39	0.42	0.37	3.7	3.5	4.5	4.9	4.1
c499	0.33	0.33	0.25	0.2	0.28	3.6	3.6	2.7	2.2	3
c880	0.28	0.28	0.29	0.29	0.29	3.1	3.3	3.3	3.2	3.2
c1355	0.34	0.25	0.29	0.15	0.26	3.7	2.8	3.3	1.6	2.8
c1908	0.37	0.36	0.38	0.38	0.37	4.1	4.1	4.2	4.2	4.1
c2670	0.43	0.4	0.4	0.39	0.4	4.6	4.3	4.3	4.1	4.3
c3540	0.42	0.46	0.42	0.4	0.43	4.5	5	4.6	4.3	4.6
c5315	0.34	0.36	0.37	0.36	0.36	3.7	3.8	4	3.9	3.8
c6288	0.25	0.26	0.31	0.34	0.29	2.8	2.9	3.5	3.8	3.3
c7552	0.39	0.38	0.37	0.37	0.38	4.2	4.1	4	4	4.1
ALU	0.3	0.27	0.29	0.21	0.27	3.3	3	3.2	2.4	3
σ_L = 1.6nm (dtd)/1.15nm (wid) and $\sigma_{V_{th}}$ = 4.7% (dtd and wid)										
c432	0.640	0.6	0.76	0.83	0.71	5.9	6.2	8.1	8.7	7.2
c499	0.630	0.58	0.42	0.35	0.5	5.7	5.4	4	3.4	4.6
c880	0.520	0.5	0.52	0.53	0.52	4.9	5	5.2	5.1	5.1
c1355	0.650	0.45	0.49	0.25	0.46	5.9	4.3	4.8	2.4	4.3
c1908	0.710	0.69	0.71	0.71	0.71	6.8	6.9	7	7.1	6.9
c2670	0.820	0.74	0.73	0.72	0.75	7.4	6.6	6.5	6.4	6.7
c3540	0.820	0.87	0.79	0.76	0.81	7.4	7.9	7.3	6.9	7.4
c5315	0.650	0.67	0.67	0.66	0.67	5.9	6.1	6.1	6	6
c6288	0.460	0.47	0.56	0.6	0.52	4.5	4.4	5.3	5.5	4.9
c7552	0.720	0.69	0.66	0.67	0.68	6.5	6.2	6	6	6.2
ALU	0.550	0.50	0.53	0.4	0.49	5.2	5.1	5.2	4	4.9
σ_L = 2 nm (dtd)/0 nm (wid) and $\sigma_{V_{th}}$ = 4.7% (dtd & wid)										
c432	1.1	1.1	1.4	1.5	1.2	8.3	9.7	13	14	11
c499	1.1	0.91	0.64	0.54	0.79	8.1	7.1	5.1	4.5	6.2
c880	0.85	0.82	0.86	0.88	0.85	6.8	7.2	7.8	7.4	7.3
c1355	1.1	0.73	0.73	0.39	0.74	8.3	5.9	6.1	3.1	5.9
c1908	1.2	1.2	1.2	1.2	1.2	9.8	10	10	10	10
c2670	1.4	1.2	1.2	1.2	1.2	10	8.8	8.7	8.6	9.1
c3540	1.4	1.5	1.3	1.2	1.3	10	11	10	9.7	10
c5315	1.1	1.1	1.1	1.1	1.1	8.2	8.5	8.3	8.1	8.3
c6288	0.76	0.75	0.88	0.91	0.83	6.4	6	7	7.1	6.6
c7552	1.2	1.1	1.1	1.1	1.1	8.8	8.3	7.9	8	8.3
ALU	0.9	0.86	0.88	0.7	0.84	7.3	7.6	7.5	6.4	7.2

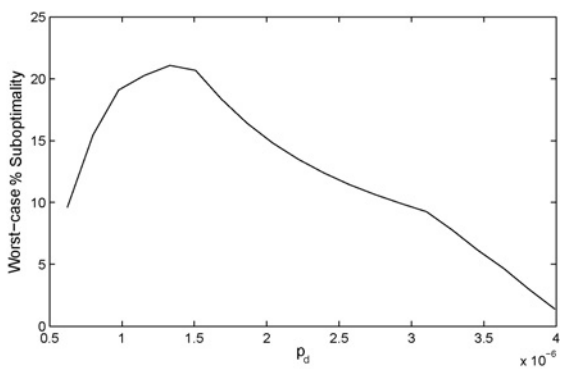


Fig. 6. Worst-case suboptimality for the example c432 v4 as a function of the assumed deterministic power value p'_d.

Another interesting thing is that adding v_{th} variations does not necessarily increase the suboptimality bound. When the length variations are small, the bounds do tend to increase. However, for large σ_L, the effect of v_{th} variations is unpredictable and small.

E. Assumption Free Bounds

The biggest weakness of the analyses above is that the deterministic optimum is not available, as the problem is too difficult to solve exactly for real circuits. In this part, we derive bounds that do not rely on an initial deterministic solution. This works by exploiting the geometry of the deterministic power measure.

We begin by assuming that only the *value* of the deterministic power measure p'_d is known. Thus, the actual deterministic optimum is one of the configurations that has a corresponding power p'_d. Out of these possible configurations, the worst-case statistical power can be found by solving the following problem:

$$\begin{array}{ll} \text{maximize} & p_s(\vec{w}) \\ \text{subject to} & p'_d = p_d(\vec{w}) \\ & w_{\min} \leq \vec{w} \leq w_{\max}. \end{array} \tag{22}$$

Denoting the optimal solution of the above as $\vec{w}_{ub}$, the worst-case suboptimality bounds can be found as follows:

$$\delta_{wc} = 100 \cdot \frac{p_s(\vec{w}_{ub}) - p_s(\vec{w}_{lb})}{p_s(\vec{w}_{lb})} \tag{23}$$

where $\vec{w}_{lb}$ is found using (17). Note that in (17), only the value p'_d and not the actual sizes of the deterministic optimum are used. As an experiment, we chose 20 equally spaced values for p'_d (between the minimum power and the maximum

TABLE VII
WORST-CASE SUBOPTIMALITY FOR $p_{m3\sigma}$

	σ_L = 1.6 nm (dtd)/1.15 nm (wid) $\sigma_{V_{th}}$ = 4.7% (dtd and wid)			
	v1 (%)	v2 (%)	v3 (%)	v4 (%)
c432	19.1	19.5	21.3	21.1
c499	18.1	15.1	10.2	8.9
c880	17.7	16.5	16.4	15.2
c1355	18.0	16.4	11.6	6.6
c1908	20.1	18.5	17.4	17.7
c2670	15.3	13.1	12.3	12.2
c3540	16.3	14.5	13.7	13.3
c5315	13.9	12.7	12.1	11.8
c6288	18.2	13.6	13.5	9.4
c7552	15.5	14.8	14.3	14.0
ALU	17.8	15.4	14.6	13.6

power), and computed the worst-case suboptimalities using (23). Running these examples for the mean + 3σ measure gives the values in Table VII, which gives values that are approximately 3x-4x the values in Table VI. Note that these numbers are only for width sizing.

It is interesting to note how the suboptimality changes as a function of p'_d. This is shown for the c432 v4 circuit in Fig. 6. At the minimum value of p'_d, there is only one possible sizing (e.g., all gates at minimum size), so the suboptimality is zero. Similarly, for the maximum value of p'_d, there is also only one size, so the suboptimality is also zero. For the remainder of the values, the suboptimality grows to a peak at a third of the way, and decreases for the remainder of the values. This implies that the suboptimality is larger for more aggressive designs, and smaller for low power designs.

F. Relating Maximum Gate Size to Suboptimality Bounds

In this section, we use (2) and (3), which model the power as a function of the gate size, to see how the bounds would change if the maximum gate sizes are increased *for all gates*.

Table VIII shows how the upper bounds in Table VI (Section V-D) increase when the gate size upper bounds are removed. In this case, the optimal deterministic power value is fixed. The values are given for the case σ_L = 1.6 nm (dtd)/1.15 nm (wid) and $\sigma_{V_{th}}$ = 4.7% (dtd and wid). When compared to Table VI, the suboptimalities for p_m increase by 0.05–0.75 percentage points, and the suboptimalities for $p_{m3\sigma}$ increase by 0.5–8.6 percentage points.

The suboptimalities increase because a better lower bound is found when larger gates are available in (17). This is because the maximum size bounds limit the sizes of the gates with the best deterministic power vs. statistical power tradeoff. As the maximum sizes increase, more of the power is used on the gates with the best tradeoffs. However, once the maximum gate size is large enough to use all of the power budget on the gates with the best tradeoff, the suboptimality will not increase any further.

The relation between suboptimality and the maximum size can be seen in Fig. 7. The increase in suboptimality from a maximum size of 4 to 8, and from 8 to 16, is approximately 3.5 percentage points. However, the increase tapers off; the increase from 16 to 32 is 1.48 percentage points and the increase from 32 to 64 is 0.01 percentage points. The suboptimalities for 64 and 128 are identical. This happens because the maximum size bound no longer plays a role in determining the lower bound.[6]

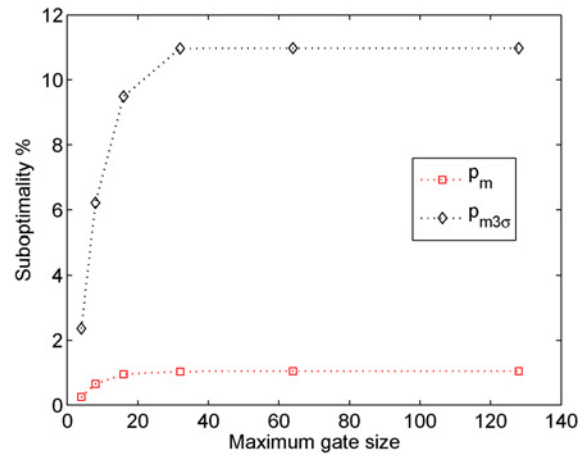


Fig. 7. Change in the suboptimality as a function of the maximum gate size in c1355 v4. Changing the maximum gate size from 2 to 128 increases the suboptimality by 8.6 percentage points.

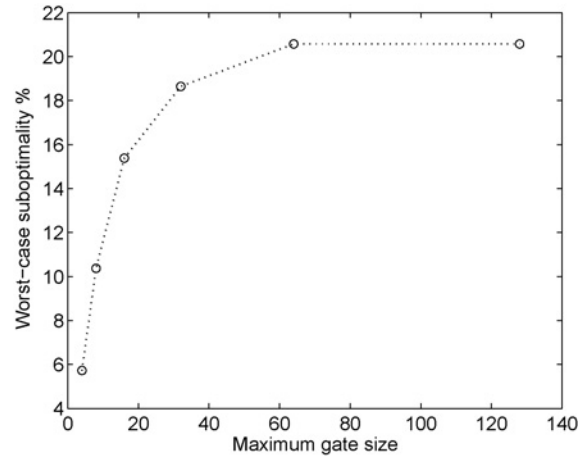


Fig. 8. Change in the worst-case suboptimality (δ_{wc}) as a function of the maximum gate size in c1355 v4 for the case in Section V-E with the same range of $p^\star_d$ as in Table VII. The worst-case suboptimality δ_{wc} reaches its limit of 20.58% for large maximum gate sizes.

Results for the "assumption free bounds" in Section V-E can also be analyzed to see how the maximum gate size affects the worst-case suboptimality δ_{wc}. Recall that in this case, the bounds are computed for a range of values for $p^\star_d$ and not just one value. Using the same range of values as used in Table VII, the suboptimality also reaches a limit as the maximum gate size increases, as in Fig. 8. The values for a maximum gate size of 64 and 128 are identical.

The explanation for this trend is similar to the case in Fig. 7. However, in this case, an upper bound is also computed in (22). This upper bound can also take advantage of the increasing maximum gate size by enabling it to use more of the power on the gates with the *worst* deterministic power vs. statistical power tradeoff.

[6]In the terminology of Section V-C, increasing the maximum sizes ceases to change the optimum in (17). This is because the optimum has sizes strictly less than the maximum gate size, and, therefore, the upper bounds do not affect the solution.

TABLE VIII

SUBOPTIMALITY (δ_{so}) FOR LEAKAGE OPTIMIZATION WITH l AND V_{th} VARIATIONS AND NO SIZE UPPER BOUNDS

σ_L = 1.6 nm (dtd)/1.15 nm (wid) and $\sigma_{V_{th}}$ = 4.7% (dtd and wid)										
	p_m					$p_{m3\sigma}$				
	v1 (%)	v2 (%)	v3 (%)	v4 (%)	Avg	v1 (%)	v2 (%)	v3 (%)	v4 (%)	Avg
c432	0.93	0.74	0.81	0.89	0.84	9.9	8.5	8.8	9.6	9.2
c499	0.92	0.88	0.93	0.92	0.91	9.8	9.8	10	10	10
c880	0.77	0.7	0.75	0.82	0.76	8.8	8.2	8.6	9.1	8.7
c1355	0.94	0.74	0.77	1	0.88	10	8.4	8.7	11	9.5
c1908	0.96	0.9	0.91	0.9	0.92	11	10	10	10	10
c2670	1.1	1	1	0.78	0.97	12	11	10	7.8	10
c3540	1.1	1.2	1.2	1.2	1.2	12	13	13	13	13
c5315	1.1	1	1	1	1	11	11	11	11	11
c6288	0.68	0.72	0.76	0.72	0.72	8.2	8	8.7	6.6	7.9
c7552	1	0.99	0.96	0.97	0.99	11	11	11	11	11
ALU	0.84	0.79	0.82	0.63	0.77	10	9.4	9.5	7.4	9.1

It is important to note that in realistic libraries, gate sizes will not go beyond 64x, with most cells being limited to 16x. As a result, numbers presented in this section are somewhat exaggerated.

G. Suboptimalities for w, v_{th}, and l Sizing

In this section, we extend these results to w, v_{th}, and l assignment. Because we do not have designs with optimal w, v_{th}, and l assignment and, hence, an optimal $z^\star$, we follow a methodology that is similar to Section V-E. Ten equally spaced values of the deterministic power are chosen ($\{p_{d_1}, ..., p_{d_{10}}\}$), where p_{d_1} is the power with all gates at the minimum power cell, and $p_{d_{10}}$ is the power with all gates at the maximum power cell. For each deterministic power, a corresponding lower bound is found (z_{lb}, as in the prior sections) and a corresponding upper bound is found (z_{ub}). The lower bound is computed by using a variation of (20) that uses a given deterministic power value p_{d_i} as follows:

$$\begin{array}{ll} \text{minimize} & p_s(\vec{z}) \\ \text{subject to} & p_{d_i} \leq p_d(\vec{z}) \\ & z_{\min} \leq \vec{z} \leq z_{\max}. \end{array} \tag{24}$$

The upper bound is computed by using a variation of (24) as a *maximization* problem as follows:

$$\begin{array}{ll} \text{maximize} & p_s(\vec{z}) \\ \text{subject to} & p_{d_i} = p_d(\vec{z}) \\ & z_{\min} \leq \vec{z} \leq z_{\max}. \end{array} \tag{25}$$

The latter problem finds the maximum statistical power configuration that has a deterministic power equal to p_{d_i}. Thus, we find the maximum statistical power that can also be an optimal deterministic power solution. Combining the two results, the maximum suboptimality is computed as follows:

$$\delta_{\max} = 100 \cdot \frac{p_s(\vec{z}^\star_{ub}) - p_s(\vec{z}_{lb})}{p_s(\vec{z}_{lb}).} \tag{26}$$

This results in a more conservative bound than in Section V-D, because a worst-case lower bound *and* a worst-case upper bound are computed as follows:

TABLE IX

SUBOPTIMALITY RATIOS $\delta_{\text{full,max}}/\delta_{w,\max}$

	σ_L = 1.6 nm (dtd) /1.15 nm (wid) $\sigma_{V_{th}}$ = 4.7% (dtd and wid)				
	v1	v2	v3	v4	Avg
c432	2.1	1.9	1.7	1.7	1.8
c499	1.7	1.8	2.9	3.6	2.5
c880	2	1.9	2	1.8	1.9
c1355	1.7	1.6	2.5	4.3	2.5
c1908	2	1.9	1.8	1.8	1.9
c2670	1.8	1.9	2	2	2
c3540	2.1	2	2.1	2.2	2.1
c5315	2.3	2.1	2.2	2.2	2.2
c6288	2	1.9	2.5	3	2.4
c7552	2	1.8	2	1.9	1.9
alu	1.8	1.7	2.1	1.4	1.7

These values are computed for width sizing ($\delta_{w,\max}$), and full w, v_{th}, and l assignment ($\delta_{\text{full,max}}$). The ratios $\delta_{\text{full,max}}/\delta_{w,\max}$ are shown in Table IX.

The ratios in Table IX can be used to relate the suboptimality bounds from Section V-D to the case of full w, v_{th}, and l assignment. This table indicates that the values will roughly double in most of the cases, but may increase up to 6.6x.

Note that in the process above, w, l, and v_{th} are not used. This is because the effect on the power is central to the bounding process, and the effect of w, l, and v_{th} on the power can be summarized by the variable z. Using the actual w, l, and v_{th} cannot be done as p_d is not quasi-concave in w, l, and v_{th}.

H. How Conservative are These Bounds?

A natural question to ask about the bounds above is, "How conservative are they?" This is because the bounds for the larger length variations are significant, and it may be useful to perform *statistical optimization* to see what the actual suboptimality is for those cases.

One indication of how conservative these bounds are comes from looking at the sizes that are used to compute the lower bound. In Fig. 9, the sizes of the deterministic optimum are plotted against the sizes that are used to compute the lower bound. The difference in the sizings is large; many of the minimum-sized gates become large gates, and vice-versa.

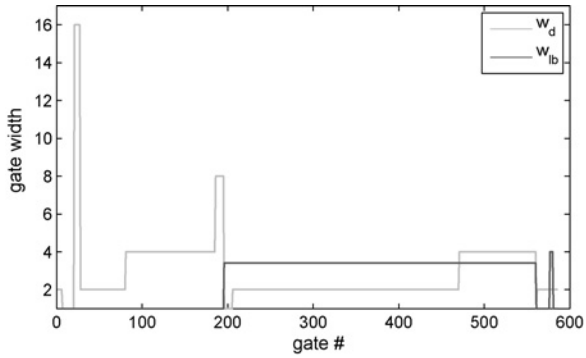


Fig. 9. Difference between the sizes that are synthesized for deterministic power (w_d), and the sizes that are used to compute the lower bound (w_{lb}). The c1355 v2 circuit is shown.

The effect of flipping the gates will generally violate timing. Gates are usually sized larger because they need to drive larger fanouts, and smaller gates are smaller because they have smaller fanouts. Thus, by flipping the sizes of these gates, small gates will need to drive larger fanouts, and there will be large gates with small fanout load.

Another intuition can be gained by re-examining Fig. 5. In both cases of (b) and (c), the actual suboptimality is 0%. However, due to the geometry of the sets and the sensitivities, the suboptimality *bounds* are very different. This indicates that the bounds may be very loose in some cases.

A final comment is related to the mathematical procedure used to create the bounds. The space was relaxed to be continuous, and the lower bound is found by finding a continuous sizing that is a bound. Adding in the discreteness constraints can only make the suboptimalities smaller.

To test the bounds, we made several small circuit examples as follows:

1) 7 gate fanout tree: 1 primary input gate, 4 primary output gates;
2) 7 gate fanin tree: 4 primary input gates, 1 primary output gate;
3) 9 gate diamond: 1 primary input gate, 1 primary output gate, with 2, 3, and 2 gates at logic depth 2, 3, 4, respectively;
4) 13 gate star: there is 1 gate in the middle at depth 3, and the fanout and fanin cones are both trees with root at this node;
5) 14 gate, 2-bit adder.

We randomly assigned the gates to these examples, and used enumeration to compute the suboptimality. A variation of σ_L = 1.6nm (dtd)/1.15nm (wid) and $\sigma_{V_{\text{th}}}$ = 4.7% (dtd & wid) is used. After running an optimization however, we found that all of these designs have suboptimality of 0%.

To find examples with suboptimalities, we randomly generated circuits according to the following rule:

1) with an N gate circuit;
2) j is a fanout of i with probability p if $j > i$, probability 0 otherwise.

We used $N = 10$ and $p = 0.5$. All the possible width combinations were enumerated, and the best designs for a series of delay targets were recorded. Fig. 10 shows one example case. The statistical and deterministic optima are the same for the majority of the cases, but they depart for a few different delay targets. In this case, the maximum suboptimality is 10.1%. This example circuit has logic depth 5, and the non-primary output gates have fan-outs of {6, 5, 5, 4, 3, 2, 1}, respectively.

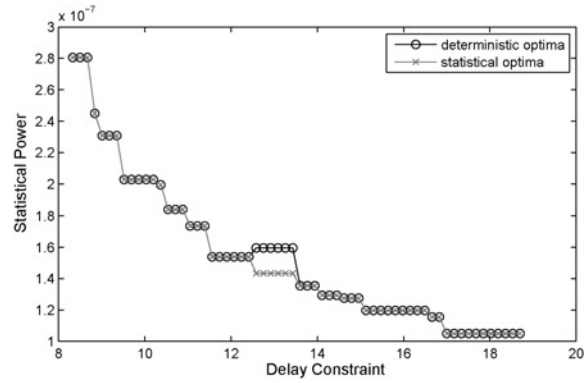


Fig. 10. Suboptimality plot for a ten gate circuit (computed by enumeration). The x-axis is the delay constraint, while the y-axis is the statistical power. The "x" denote statistical minima and the "o" denote deterministic optima. When the black and gray lines diverge, there is a suboptimality gap, which is the difference between the two lines. The maximum gap in this plot is 10.1%. Note that for the majority of the delay constraints, the deterministic and statistical optima are the same.

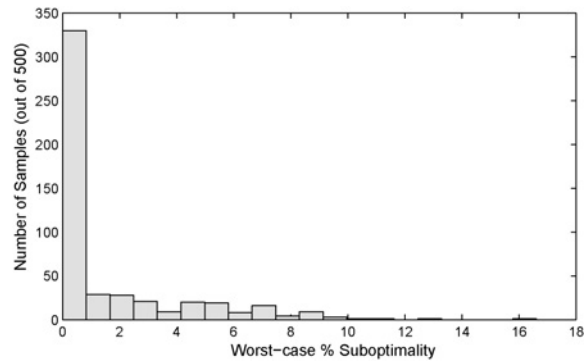


Fig. 11. Histogram of the suboptimalities for 500 randomly generated circuits. The majority of the examples have suboptimality of 0%.

We generated 500 of these random circuits, and computed the worst-case suboptimality for each circuit across all delay targets. Most of the circuits that were generated have 0% suboptimality (see the histogram in Fig. 11), however the worst-case suboptimality was 16.6% which correlates with the numbers in Table VII. This shows that the suboptimality is likely to be 0%, although the worst-case numbers may be significant.

VI. Bridging the Suboptimality Gap

The suboptimality bounds for large l variations cause some reason for concern. Some of the bounds are over 10%, which is a significant amount that is too large to ignore. In this section we present ways to the suboptimality gap using simpler measures.

There is a significant cost to using the $p_{m3\sigma}(\cdot)$ measure. It is significantly more complicated than the other power measures.

TABLE X

WORST-CASE SUBOPTIMALITY (δ_{wc}) OF p_{approx} AS AN ESTIMATE OF $p_{m3\sigma}$ FOR A MAXIMUM SIZE OF 32, AND w, l, v_{th} ASSIGNMENT

σ_L = 1.6 nm (dtd)/1.15 nm (wid) $\sigma_{V_{\text{th}}}$ = 4.7% (dtd and wid)					
	v1 (%)	v2 (%)	v3 (%)	v4 (%)	Avg
c432	2.2	1.9	1.6	1.4	1.8
c499	2.0	2.1	2.0	1.4	1.9
c880	2.1	2	2	1.9	2.0
c1355	2.1	2.2	1.7	1.6	1.9
c1908	2.2	2.2	2.2	2.2	2.2
c2670	1.9	1.7	1.6	1.6	1.7
c3540	2.1	2.2	2.1	2.1	2.1
c5315	1.8	1.7	1.7	1.6	1.7
c6288	2.1	3.1	2.7	1.4	2.3
c7552	1.7	1.6	1.6	1.6	1.6
ALU	2.2	3.5	3.0	3.5	3.0

While the deterministic ($p_d(\cdot)$) or the mean ($p_m(\cdot)$) power measures are linear in z, the $p_{m3\sigma}(\cdot)$ is nonlinear in z (although it is convex in z). Linear measures have the advantage that the total power is the sum of the individual powers. Thus, these measures can be used in existing optimization methods by replacing the power values in the library files with the statistical power values.

Fortunately, there is a simple linear approximation that we can use to bridge the suboptimality gap. The idea is to use the variation information to construct a linear measure that has a low suboptimality gap. This would result in a measure that is a provably good approximation to the full statistical optimization, and can be used by replacing the deterministic power values by these approximate values. We define this approximation to the $p_{m3\sigma}$ measure as follows:

$$p_{\text{approx}} = \sum \kappa_i \exp\left(\lambda_{\text{w}(i)}\sigma_{\mathbf{L}_{\text{w}(i)}} + \lambda_{\text{d}(i)}\sigma_{\mathbf{L}_{\text{d}}} + \ldots \sqrt{2}(\tau_{\text{w}(i)}\sigma_{\mathbf{v}_{\text{th}_{\text{w}(i)}}} + \tau_{\text{d}(i)}\sigma_{\mathbf{v}_{\text{th}_{\text{d}}}})/2\right). \quad (27)$$

The derivation of this approximation is purely empirical; this measure was found to give very small suboptimality bounds. It can be interpreted as using the 1-σ value of the random variables $\mathbf{v}_{\text{th}_{\text{w}(i)}}$, $\mathbf{v}_{\text{th}_{\text{d}}}$, $\mathbf{L}_{\text{w}(i)}$, and $\mathbf{L}_{\text{d}}$. The $\sqrt{2}/2$ factor used to account for intra-die cancellation. This measure performs better than the p_m measure and it is also better than the 3-σ value of the random variables.

Worst-case suboptimality bounds for this approximation are given in Table X for full w, v_{th}, and l assignment (as in Section V-G), 1–32x sized gates (as in Section V-F), and the same range for $p_d^\star$ as used in Table VII. This measure is very effective, as it reduces the suboptimality gap to 3.5% and less. This is around an 80% reduction in the suboptimalities. This indicates that although the $p_{m3\sigma}$ measure may be different from the deterministic measure, the sensitivities can be well approximated by a linear measure.

In contrast p_m is not as good of a linear approximation to $p_{m3\sigma}$. The resulting suboptimality bounds give about a 10% reduction in the suboptimalities.

VII. CONCLUSION

In this paper, we compared deterministic solutions of sizing problems with statistical solutions of sizing problems. The rankings of the solutions coincide very well for the mean power measure, indicating that good deterministic solutions will be equally good mean power solutions. However, the other measures have significant discrepancy in their rank. These findings were substantiated by computing the worst-case bounds on the suboptimality gap. The bounds are always insignificant for the mean-power measure p_m, but they may become significant for the $p_{m3\sigma}$ measure when the inter-die component of the length variations are large. As a way to bridge the suboptimality gap, we presented a proxy measure for the $p_{m3\sigma}$ measure, which is an excellent approximation, and has an insignificant suboptimality.

ACKNOWLEDGMENT

We are very thankful to Dr. S. Shah for his insights on statistical power optimization.

REFERENCES

[1] J. Cong, P. Gupta, and J. Lee, "On the futility of statistical power optimization," in *Proc. Asia South Pacific Design Autom. Conf.*, 2009, pp. 167–172.
[2] S. Nassif, A. Strojwas, and S. Director, "A methodology for worst-case analysis of integrated circuits," *IEEE TCAD Integr. Circuits Syst.*, vol. 5, no. 1, pp. 104–113, Jan. 1986.
[3] C. Visweswariah, "Death, taxes and failing chips," in *Proc. Design Automat. Conf.*, 2003, pp. 343–347.
[4] M. Mani and M. Orshansky, "A new statistical optimization algorithm for gate sizing," in *Proc. IEEE Int. Conf. Comput. Design*, 2004, pp. 272–277.
[5] D. Patil, S. Yun, S. Kim, A. Cheung, M. Horowitz, and S. Boyd, "A new method for design of robust digital circuits," in *Proc. ISQED*, 2005, pp. 676–681.
[6] J. Singh, V. Nookala, Z. Luo, and S. Sapatnekar, "Robust gate sizing by geometric programming," in *Proc. Design Automat. Conf.*, 2005, pp. 315–320.
[7] K. Chopra, S. Shah, A. Srivastava, D. Blaauw, and D. Sylvester, "Parametric yield maximization using gate sizing based on efficient statistical power and delay gradient computation," in *Proc. Int. Conf. Comput.-Aided Design*, 2005, pp. 1023–1028.
[8] M. Guthaus, N. Venkateswaran, C. Visweswariah, and V. Zolotov, "Gate sizing using incremental parameterized statistical timing analysis," in *Proc. Int. Conf. Comput.-Aided Design*, 2005, pp. 1029–1036.
[9] D. Sinha, N. Shenoy, and H. Zhou, "Statistical gate sizing for timing yield optimization," in *Proc. Int. Conf. Comput.-Aided Design*, 2005, pp. 1037–1041.
[10] A. Davoodi and A. Srivastava, "Variability driven gate sizing for binning yield optimization," in *Proc. Design Automat. Conf.*, 2006, pp. 959–964.
[11] A. Srivastava, D. Sylvester, and D. Blaauw, "Statistical optimization of leakage power considering process variations using dual-Vth and sizing," in *Proc. Design Automat. Conf.*, 2004, pp. 773–778.
[12] M. Mani, A. Devgan, and M. Orshansky, "An efficient algorithm for statistical minimization of total power under timing yield constraints," in *Proc. Design Automat. Conf.*, 2005, pp. 309–314.
[13] S. Bhardwaj and S. B. K. Vrudhula, "Leakage minimization of nano-scale circuits in the presence of systematic and random variations," in *Proc. Design Automat. Conf.*, 2005, pp. 541–546.
[14] X. Ye, Y. Zhan, and P. Li, "Statistical leakage power minimization using fast equi-slack shell based optimization," in *Proc. Design Automat. Conf.*, 2007, pp. 853–858.
[15] J. Cong, J. Lee, and L. Vandenberghe, "Robust gate sizing via mean excess delay minimization," in *Proc. ISPD*, 2008, pp. 10–14.
[16] F. N. Najm, "On the need for statistical timing analysis," in *Proc. Design Automat. Conf.*, 2005, pp. 764–765.

[17] M. Guthaus, N. Venkateswaran, V. Zolotov, D. Sylvester, and R. Brown, "Optimization objectives and models of variation for statistical gate sizing," in *Proc. Great Lakes Symp. VLSI*, 2005, pp. 313–316.
[18] S. M. Burns, M. Ketkar, N. Menezes, K. A. Bowman, J. W. Tschanz, and V. De, "Comparative analysis of conventional and statistical design techniques," in *Proc. Design Automat. Conf.*, 2007, pp. 238–243.
[19] *Nangate Open Cell Library v1.2* [Online]. Available: http://www.si2.org/openeda.si2.org/projects/nangatelib
[20] R. Rao, A. Srivastava, D. Blaauw, and D. Sylvester, "Statistical analysis of subthreshold leakage current for VLSI circuits," *IEEE Trans. VLSI Syst.*, vol. 12, no. 2, pp. 131–139, Feb. 2004.
[21] R. Usselman. (2000). "Mini-RISC-1 ALU." *OpenCores* [Online]. Available: http://www.opencores.org
[22] Cadence Design Systems, Inc. (2007). *Encounter RTL Compiler 6.20* [Online]. Available: http://www.cadence.com
[23] The MathWorks. (2008). *MATLAB R2008b* [Online]. Available: http://www.mathworks.com
[24] W. Li, "Strongly NP-hard discrete gate sizing problems," in *Proc. IEEE Int. Conf. Comput. Design: VLSI Comput. Process.*, Oct. 1993, pp. 468–471.
[25] M. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, nos. 1–2, p. 81, 1938.

Jason Cong (S'88–M'90–SM'96–F'00) received the B.S. degree in computer science from Peking University, Beijing, China, in 1985, and the M.S. and Ph.D. degrees in computer science from the University of Illinois at Urbana-Champaign (UIUC), Urbana, in 1987 and 1990, respectively.

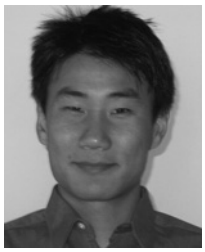
He is currently a Chancellor's Professor with the Department of Computer Science, University of California, Los Angeles (UCLA), the Director of Center for Domain-Specific Computing (funded by an NSF Expeditions in Computing Award), a Co-Director of UCLA/Peking University Joint Research Institute in Science and Engineering and a Co-Director of the VLSI CAD Laboratory, Department of Computer Science, UCLA. He also served as the Department Chair from 2005 to 2008. He has served on the technical advisory boards of a number of EDA and silicon IP companies, including Atrenta, eASIC, Get2Chip, Magma Design Automation, and Ultima Interconnect Technologies. He was the Founder and President of Aplus Design Technologies, Inc., Los Angeles, until it was acquired by Magma Design Automation, San Jose, CA, in 2003. He currently serves as the Chief Technology Advisor and Chairman of the Board of AutoESL Design Technologies, Inc., Cupertino, CA. He has graduated 26 Ph.D. students. A number of them are now faculty members in major research universities, including the Georgia Institute of Technology, Atlanta, Purdue University, West Lafayette, IN, the State University of New York at Binghamton, Binghamton, NY, the UCLA, UIUC, and the University of Texas at Austin, Austin. Others are taking key research and development or management positions in major EDA/computer/semiconductor companies or founding high-tech startups. His current research interests include computer-aided design of VLSI circuits and systems, design and synthesis of system-on-a-chip, programmable systems, novel computer architectures, nanosystems, and highly scalable algorithms. He has published over 300 research papers and led over 30 research projects in these areas.

Dr. Cong has received a number of awards and recognitions, including the Ross J. Martin Award for Excellence in Research from UIUC in 1989, the NSF Young Investigator Award in 1993, the Northrop Outstanding Junior Faculty Research Award from UCLA in 1993, the ACM/SIGDA Meritorious Service Award in 1998, and the SRC Technical Excellence Award in 2000. He received five Best Paper Awards selected for the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN in 1995, the International Symposium on Physical Design in 2005, the ACM Transaction on Design Automation of Electronic Systems in 2005, the International Symposium on High Performance Computer Architecture in 2008, and the IEEE Symposium on Application Specific Processors in 2009. He was elected a Fellow of ACM in 2008.

Puneet Gupta (S'02–M'08) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology Delhi, New Delhi, India, in 2000, and the Ph.D. degree from the University of California, San Diego, in 2007.

He is currently a Faculty Member with the Electrical Engineering Department, University of California, Los Angeles. He is the Co-Founder of Blaze DFM Inc. (acquired by Tela Inc.) in 2004 and served as its Product Architect until 2007. He has authored over 60 papers, eight U.S. patents, and a book chapter. His current research interests include building high-value bridges between physical design and semiconductor manufacturing for lowered cost, increased yield, and improved predictability of integrated circuits.

Dr. Gupta has been a recipient of the NSF Career Award, the ACM/SIGDA Outstanding New Faculty Award, the IBM Ph.D. Fellowship, and the European Design Automation Association Outstanding Dissertation Award. He has given tutorial talks at DAC, ICCAD, the International VLSI Design Conference, and the SPIE Advanced Lithography Symposium. He has served on the technical program committees of DAC, ICCAD, ASPDAC, ISQED, ICCD, SLIP, and VLSI Design. He served as the Program Chair of the IEEE DFM&Y Workshop in 2010.



John Lee (S'08) received the B.A. and B.S. degrees from the University of California, Berkeley, in 2003, and the M.S. degree from the University of California, Los Angeles, where he is currently pursuing the Ph.D. degree in electrical engineering.

His current research interests include optimization and applied mathematics, with applications in physical design.